UNIVERSITY OF CALIFORNIA

Santa Barbara

Observing Snow from Space: Snow/Cloud Discrimination and Opportunities in Water Supply Forecasting

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Environmental Science and Management

by

Timbo Conrad Stillinger

Committee in charge:

Professor Jeff Dozier, Chair

Professor Dar Roberts

Professor Chris Costello

Professor Roger Bales, University of California Merced

December 2019

The dissertation of Timbo Conrad Stillinger is approved.

_____

Roger Bales

_____

Dar Roberts

_____

Chris Costello

_____

Jeff Dozier, Committee Chair

October 2019

Observing Snow from Space: Snow/Cloud Discrimination & Opportunities in Water Supply Forecasting

By

Timbo Conrad Stillinger

ACKNOWLEDGEMENTS

This dissertation would not have been possible without Jeff Dozier. His style of mentorship, which encouraged independence and fostered my confidence to tackle new and complex problems from scratch, formed the scientist that I am today. I am forever grateful for his support and enabling my convenient excuse to visit Mammoth Lakes for "work".

To Dar Roberts, for teaching me nuances of passive remote sensing and encouraging my conceptual understanding of the physical processes we measure; to Chris Costello's enthusiasm behind my ideas and willingness to let me pretend to be an economist for a chapter; to Roger Bales mentorship in the process of crafting a scientific paper and help with understanding snow hydrology, water operations, and water policy in California I am very thankful.

This thesis would not have been possible without the continued support of my wife, Maggie Crawford. While faster than me in the mountains and at graduate school, her positive encouragement and adventurous distractions significantly contributed to my success. A big thank you as well to my parents, whom have always encouraged my curiosity. This experience would not have been the same without Ty Brandt, my academic brother in this long journey. I'll miss the coffee breaks and British humor.

My formal education since kindergarten has been in California public schools. I am grateful to all my teachers and to the taxpayers of California who provided the financial support for my education. And finally to Muir, the best dog ever, for always being there to work out ideas with me on our walks.

VITA OF TIMBO CONRAD STILLINGER

October 2019

EDUCATION

Bachelor of Science in Molecular Environmental Biology, University of California, Berkeley, May 2010
Master of Environmental Science and Management, University of California, Santa Barbara, June 2014
Doctor of Philosophy in Environmental Science and Management, University of California, Santa Barbara, October 2019 (expected)

PROFESSIONAL EMPLOYMENT

2011-12: Associate Biologist, San Elijo Lagoon Conservancy, Encinitas, California
2013-19: Teaching Assistant, Bren School of Environmental Science and Management, University of California, Santa Barbara
2013-19: Graduate Student Researcher, Earth Research Institute, University of California, Santa Barbara

PUBLICATIONS

Palazzo, J., Liu, O. R., Stillinger, T., Song, R., Wang, Y., Hiroyasu, E. H. T., Zenteno, J., Anderson, S., and Tague, C. (2017), Urban responses to restrictive conservation policy during drought, Water Resour. Res., 53, 4459– 4475, doi: 10.1002/2016WR020136.
Stillinger, T., Roberts, D.A., Collar, N.M., and Dozier, J. (2019). Cloud masking for Landsat 8 and MODIS Terra over snow-covered terrain: Error analysis and spectral similarity between snow and cloud. Water Resour. Res., 55, 6169-6184, doi: 10.1029/2019WR024932.
Stillinger, T., & Collar, N.M. (2019). Snow-Cloud Validation Masks for Multispectral Satellite Data. Version 1. doi: 10.5281/zenodo.3240937

AWARDS

Graduate Student Association Excellence in Teaching Award in Science, Technology, Engineering, and Mathematics, 2017
Bren School of Environmental Science & Management Outstanding Teaching Assistant Award, 2019.

FIELDS OF STUDY

Remote Sensing, Snow Hydrology, Forecasting

ABSTRACT

Observing Snow from Space: Snow/Cloud Discrimination and Opportunities in Water Supply Forecasting

By

Timbo Conrad Stillinger

Seasonal snow and ice support important ecosystems and comprise the main water supply for nearly two billion people, yet forecasts of snowmelt runoff either do not exist or occasionally significantly under- or over-predict flows. Mapping snow cover and albedo from space has enabled a paradigm shift in our ability to observe processes that occur in the mountains. However, clouds commonly obstruct the surface in visible and infrared spectra, and existing operational cloud masks consistently misclassify some snow as clouds and some clouds as snow. This work aims to both advance our ability to utilize satellites for observing the mountain snowpack and to improve our understanding of the potential operational benefits of better forecasts from these tools.

In Chapter 1, I assess the errors in the cloud masks over snow-covered, mid-latitude mountains for the Landsat 8 Operational Line Imager (OLI) and for the Moderate-Resolution Imaging Spectroradiometer (MODIS) on the Terra satellite: CFMask for Landsat 8 and the cloud mask that ships with the MOD09GA and MYD09GA surface reflectance products from MODIS. The overall precision and recall of CFMask

over snow-covered terrain are 0.70 and 0.86; the MOD09GA cloud mask precision and recall are 0.17 and 0.72. I find a plausible reason for poorer performance of cloud masks over snow lies in the potential similarity between multispectral signatures of snow and cloud pixels in three situations: (1) Snow at high elevation is bright enough in the "cirrus" bands (Landsat band 9 or MODIS band 26) to be classified as cirrus. (2) Reflectances of "dark" clouds in shortwave infrared (SWIR) bands are bracketed by snow spectra in those wavelengths. (3) Snow as part of a fractional mixture in a pixel with soils sometimes produces "bright SWIR" pixels that look like clouds.

In Chapter 2, I develop a new method for snow/cloud discrimination that relies on textural and spatial features alongside the spectra to identify clouds and their optical properties. I extend the applicability of superpixels to multispectral satellite data by generating superpixels from an 8-D space of the eight 30m Landsat 8 optical bands by mapping the spectral angle between superpixel mean spectra and each component pixel. A Gabor filter bank on Landsat 8 panchromatic data accurately discriminates mountain snow that is spectrally similar to clouds. Clouds and shadows are also identified by using the expected topographic shading alongside spectral tests to separate clouds, terrain shadow, cloud shadow, and shaded snow. Histograms of oriented gradients identify mountainous terrain and obstruction of terrain illumination patterns by clouds. A radiative transfer model assigns optical properties to cloud mask pixels. This new method improves the precision of Landsat 8 cloud masks by 31% (from 70% to 91%) over mid-latitude snow covered mountains compared to the operational product.

To understand the opportunities for remote sensing of snow for water operations, in Chapter 3 I use thirty-four years of water data in fourteen California Sierra Nevada basins to examine the association between water management decisions and snowmelt runoff forecasts, their uncertainty, and error. The findings indicate that forecasts are positively associated with releases, but increased uncertainty in a forecast is negatively associated with releases. Reduction in spring runoff forecast uncertainty is possible from remote sensing tools and would enable additional uses of storing runoff from April through July.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

Seasonal snow and ice comprise the main supply of water for 2 billion people around the world. Yet, the rugged environment of the mountains combined with the spatial heterogeneity of snow accumulation and melt makes forecasting water supplies from seasonal snow a challenge. For some of the populations most dependent on seasonal snow and ice, no in situ measurements or forecasts of their water supply exist. Satellites provide the most consistent and reliable measurements of snow and ice globally with estimates of snow-covered area, grain size, and concentration of light-absorbing particulates creating global potential to improve existing operational forecasts and enable forecasting of water supplies for regions with little to no in situ forecasting.

A main hurdle to consistent, accurate estimates of snow properties from space is the accurate and automatic identification of clouds in satellite images. Clouds both cover snow from the view of sensors in the visible and infrared and impact the energy balance of the snowpack as it accumulates and melts. This thesis documents three research investigations focused around measuring snow and forecasting snowmelt runoff: (i) an analysis of root causes for difficulty in snow and cloud spectral discrimination in multispectral satellite data; (ii) a textural and spatial approach to snow and cloud discrimination that clarifies the discrimination issues; (iii) an estimate of the effect of uncertainty in in-situ operational snowmelt runoff forecasts on reservoir operations as a conservative estimate of the opportunity for improvement in water management with satellite estimates of snow properties.

In Chapter 1, I assess the performance of exiting cloud masks over snow covered mountains in the Landsat 8 Operational Land Imager (OLI) and the Moderate Resolution Imaging Spectroradiometer (MODIS). Spectral libraries of clouds and snow from OLI are then used to examine the scenarios when snow and cloud can have similar spectral characteristics in the band passes of multispectral sensors. Lastly, to prove that spectral discrimination is very challenging for some snow and clouds, measured snow and cloud spectra are successfully modeled as both snow and cloud spectra with no ability to discriminate in Landsat 8 band passes based on the common differences between snow and cloud. Models of snow and clouds with the same physical properties at the spectral resolution of an imaging spectrometer show that at this higher resolution, discrimination based on spectra would still be possible for these difficult pixels.

In Chapter 2, I use what was learned about the spectral similarities and differences for various types of snow and clouds to develop an approach to cloud masking over snow that incorporates spatial and textural features alongside spectra for accurate cloud identification. An accurate cloud mask enables per-pixel estimates of snow cover, grain size, and contaminate concentration. Insights from the methods used by human analysts to create validation cloud masks are the foundation for the new textural and spatial approaches. Gabor filters, a proven model for low-level vision response to textures in mammalian cortical cells, are applied to the panchromatic band of OLI with specific tunings that create distinct contrast between clouds and snow covered forests, alpine snowpack boundaries, and low fractional snow cover, three of the most difficult snow classes to separate from cloud spectra based on per pixel

spectral tests. Superpixels are then used to segment the image into groups of pixels based on regional low-level image characteristics and replace pixels as the spatial resolution of the algorithm; this greatly reduces the image size but not the information. Superpixels are identified in OLI images by mapping the spectral angle between generated superpixels that weight both the Euclidian distance between pixels and the spectral angle similarity in 8-D space within local regions of the image. This is the first known application of traditional 3D superpixels to both a spectral angle based classifier and more than 3 dimensions. This approach is unique from other modern approaches to cloud masking in OLI in that no machine learning or training data are needed to create accurate cloud masks and some optical properties of the clouds are estimated as well.

Chapter 3 assesses the effect of uncertainty in seasonal water supply forecasts on reservoir management in the Sierra Nevada of California. I find that if forecast uncertainty could be reduced, possibly by incorporating additional remotely sensed measurements of the Sierra Nevada snowpack, the improvement in official water supply forecasts would have a significant impact on water management in the state. I find on average forecast uncertainty decreases the volume released from April-July by 250 acre-feet for each additional 1000 acre-feet of forecast uncertainty. Knowing that satellite based remote sensing of snow enables global measurements of snow properties and that improved cloud masks will enable more reliable estimates, Chapter 3 estimates how much room for improvement there is in reservoir management with a panel data analysis of forecasts and operations for one of the largest water systems in the world that relies on one of the most advanced long term in-situ networks used for

operational forecasting. Historical records of snowmelt runoff forecast, measured runoff, and reservoir releases are used to determine the effect of forecast uncertainty on basin releases.

1 Cloud Masking for Landsat 8 and MODIS Terra over Snow-Covered Terrain: Error Analysis and Spectral Similarity between Snow and Cloud

1.1 Introduction

Mapping snow cover and albedo from space is our most convenient and consistent way to observe processes that occur in the world's mountain ranges. The expanse of the mountain environment and unsettled nature of mountain weather encumber in situ and airborne data collection. Even though satellites view the mountains safely removed from these difficulties, clouds obscure parts of Earth's surface from the view of sensors in solar and infrared spectra. Knowing which pixels are snow-covered in cloudy scenes is necessary to map the daily, weekly, and monthly changes in spatial extent, grain size, and prevalence of light-absorbing particles of the snowpack (Dozier & Painter, 2004; Zhang et al., 2019). Accurate snow cover and albedo from remote sensing data are needed for estimating rates of snowmelt runoff (Painter et al., 2018), reconstructing the snow water equivalent (Bair et al., 2019), or understanding interannual variability and trends (Skiles et al., 2012). A serious, yet tractable, challenge to the consistent delivery of timely high-quality information about mountain snow from spaceborne multispectral sensors is the identification of clouds in images of snow-covered terrain. This study examines the accuracy of Landsat 8 and MODIS cloud masks over mid-latitude snow-covered mountains and highlights the difficulties for cloud masking in mountain hydrology.

Many clouds are easily identified from their visible and infrared spectra. Clouds are brighter and colder than most land surfaces except snow and ice, so for many users of remote sensing data, existing cloud masks can be trusted to identify clouds (Foga et al., 2017; Zhu et al., 2015). Since clouds tend to move, and sensors pass over the same location repeatedly, users can analyze the land surface without significant errors from cloudy pixels by using clear-sky pixels from the clouds' neighborhood from a single scene or amalgamate clear-sky pixels from multiple overpasses.

Most clouds are spectrally distinct from snow. The smaller sizes of the scattering droplets or ice crystals in clouds, compared to sizes of snow grains, cause the main differences between the spectral reflectance of snow and clouds (Dozier, 1989), with snow density having little effect on the snow's spectral reflectance (Bohren & Beschta, 1979; Warren, 1982). In the visible spectral region where ice and water are transparent, clouds and snow are both bright. In the shortwave-infrared region, ice and water are moderately absorptive, so the smaller scattering elements in clouds, whether water droplets or ice crystals, usually make them brighter than snow (Warren, 1982).

Clouds frequently obstruct mountainous surfaces, and consistent cloud cover can block clear-sky acquisitions of some areas for long periods depending on a satellite's orbit. Despite distinct radiative properties of clouds and snow, accurate, dependable cloud masks do not yet exist for use in large-scale scientific data analysis of Earth's snow- and ice-covered regions where clouds often persist. Both snow and clouds are bright, white, cold objects composed of water in either the frozen or the liquid state, and these similarities can make discrimination between snow and cloud from spaceborne multispectral sensors difficult. Persistently misclassified snow as

cloud or cloud as snow affect models of snow accumulation and melt that rely on accurate snow-covered area and albedo and local radiative conditions for energy balance modeling (Bair et al., 2016). Cloud masking algorithms perform the worst in snow-covered mountain ranges where some clouds are identified as snow and some snow is identified as cloud (Hall & Riggs, 2007; Irish et al., 2006; Zhu & Woodcock, 2012).

Few analyses explicitly test these cloud masking algorithms for their performance over snow-covered terrain, especially in mountains. For example, the snow and ice regions chosen to validate the Landsat 8 Cloud Cover Assessment (U.S. Geological Survey, 2016) exclude mid-latitude mountainous regions and instead focus on the polar regions where validation imagery consists mostly of snow and ice with little variability in topography. Therefore, snowy mountainous regions are under-sampled and binned into disparate categories when global cloud analyses create a validation dataset that covers the entire globe (Foga et al., 2017; Hughes & Hayes, 2014). Mountains cover about a quarter of Earth's land surface, provide seasonal water supply for nearly two billion people and include important ecosystems (Mankin et al., 2015; Wester et al., 2019). In North America, for example, mountains cover 24% of the area but contain ~60% of the snow (Wrzesien et al., 2018). Assessing remote sensing tools based on their global performance undervalues the importance of these mountainous regions to environmental management. Thus, the importance of mountains in supplying water and the spatiotemporal variability of seasonal snow heightens the need for better cloud masks. Since images of mountainous regions are often just partly cloudy during the snow season, usually some clear-sky pixels exist in

most scenes. These clear-sky pixels provide information for analyses only if they can be reliably discriminated from cloudy pixels.

1.2 Methods

1.2.1 Cloud Masks Assessed

We assess the accuracy of the operational cloud masks that ship with data from the Landsat 8 Operational Land Imager (OLI: Foga et al., 2017; Roy et al., 2014) and the Moderate-Resolution Imaging Spectroradiometer (MODIS: Justice et al., 1998; Platnick et al., 2003). Users often start the data-filtering step of remote sensing analyses with these products. The new Landsat standard, CFMask, now ships in the BQA file of all Landsat 8 data. The cloud mask for the MOD09GA and MYD09GA products originates from the MOD35 cloud masking code and also forms the basis for approaches that utilize additional information from time series (Friedl et al., 2002; Platnick et al., 2003). Our validation of the MODIS cloud mask utilizes data from Landsat 8 (section 1.2.3.1).

Snow varies at a finer spatial scale than clouds. In mountain ranges worldwide, reflectance signatures of nearly every MODIS snow-covered pixel and 93% of Landsat snow-covered pixels indicate the pixel includes other land covers (Selkowitz et al., 2014). In contrast, only around 15% of global cloud cover are from clouds smaller than 10 km^2, where most of these "small" clouds are still much larger than Landsat and MODIS pixels (Wood & Field, 2011). For this analysis, we assume clouds are bigger than the sensor's pixel hence no fractional cloud-covered pixels exist, whereas snow varies at a finer scale and fractional snow-covered pixels are plentiful.

For reference, Table 1 shows the band passes and resolutions of the Landsat OLI (bands 1-9) and the "land" bands from MODIS (bands 1-7 and the cirrus band 26).

1.2.2 Validation Datasets

We used two datasets of manually classified Landsat 8 OLI data, 13 scenes that we analyzed and 13 from the U.S. Geological Survey (2016), to assess the performance of the Landsat 8 and MODIS cloud masks. Figure 1 shows the locations of the validation scenes. Landsat 8 validation datasets are a valuable assessment tool for MODIS because acquisitions from the Landsat 8 and Terra satellites occur within 20 minutes of each other on similar near-polar descending orbits. Manual classification of pixels is more difficult on the coarser 250-1000 m resolution MODIS data compared to the 30 m OLI data. The close acquisition times of the two satellites minimizes misinterpretation caused by cloud movement.

Any manual classification approach that relies on Landsat imagery for creating validation data will misclassify some pixels. Foga et al. (2017) estimate the difference between individual analysts at 7%. Thin clouds that do not have distinguishing reflectance in OLI band 9 and some small clouds fewer than ~10 pixels are difficult for a human to classify. Snow pixels with fractional snow-covered area below 50% can be difficult to identify. Along the borders of a snowpack, some of the pixels with lower snow-covered area are mistakenly excluded. Some snow-free pixels may mistakenly be included because exactly tracing the boundary of every snowpack and cloud in 30 m spatial data is impossible.

Within the 13 scenes that we selected, we manually masked snow, cloud, and snow-free pixels. Using USGS Earth Explorer (https://earthexplorer.usgs.gov/), we previewed false-color scenes over the Western U.S. and the Himalaya to find images that included a variety of cloud types and areas of cloud cover over snow, forest, and bare ground. Scenes with complete cloud cover or no clouds were excluded. Within chosen scenes, we randomly located regions that contained cirrus clouds, thin clouds, and opaque clouds co-located with snow-covered mountains. Pixels were then manually classified as cloud, snow, or "neither," i.e. clear-sky and snow-free, containing either land or open water. Reference layers were created in ArcGIS using various techniques to enable visualization of each cover type. Snow pixels were identified using natural color images of Landsat 8 OLI bands 4, 3, and 2, false color images of bands 6, 5, and 4, and a mask of the Normalized Difference Snow Index (NDSI: Dozier, 1989) from bands 3 and 6. Opaque clouds were identified using the same natural color bands 4, 3 and 2 along with greyscale images of bands 2, 4 and 5 and false color images of bands 2, 4 and 5. Images from the cirrus band (band 9) were used to identify clouds high enough in the atmosphere where minimal water vapor existed above them. Pixels that were manually identified as more than one cover type were excluded from the reference masks.

The second validation dataset of 13 scenes came from the 2014 USGS SPARCS manual classification dataset (U.S. Geological Survey, 2016) of Landsat 8 OLI sub-scenes of 1000×1000 pixels, with each pixel identified as snow, cloud, cloud shadow, water, land, or bad data. Of 80 scenes in this dataset, 13 included significant snow within the manual masks. We used these 13 scenes and the 13 that we classified to

assess the performance of the operational CFMask in areas of snow cover. We categorized water and land pixels as "neither" given our focus on the ability to distinguish clouds from snow. We excluded the pixels labeled as cloud shadows from the analysis because that class aggregates both shaded snow and shaded vegetation and soil.

Both sets of validation data were generated on Pre-Collection Landsat data that were then aligned to Collection 1 Landsat scenes. The manually identified pixels were compared to the pixel classifications in the BQA file of the Collection 1 data (U.S. Geological Survey, 2017), specifically the CFMask bits for cloud and cirrus, which were merged into a single binary cloud mask. Most users will employ this default CFMask, which Foga et al. (2017) determined to be the best cloud mask for Landsat 8 data.

1.2.3 Error Assessment Metrics

We statistically evaluated the performance of the extant Landsat OLI snow and cloud algorithms and MODIS cloud algorithm against the manual masks from OLI. We compared the algorithms' classifications for each class of interest (snow, cloud, neither) for each pixel to the manual classification. For each class of interest (snow, cloud, neither), we categorized each pixel as:

True Positive – TP: Both the automated mask and analysts classify the pixel as the class of interest.

True Negative – TN: Both the automated mask and analysts classify the pixel as not the class of interest. The classification of the analyst and the automated mask do not have to match.

False Negative – FN: Analysts classify the pixel as the class of interest, and the automated mask classifies it differently.

False Positive – FP: Automated mask classifies a pixel as the class of interest, and the analysts classify it differently.

For each scene and for all scenes aggregated together, the total counts of each category (TP, FP, TN, FP) for each class of interest (snow, cloud, neither) are totaled. The following metrics are used in the error assessment of the algorithms (Olson & Delen, 2008):

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{3}$$

Precision represents the probability that a pixel identified as a class is indeed that class. *Recall* represents the probability of detection of a pixel of the class of interest. The *F* score balances the two metrics. Some error analyses use *Omission* (missing the class of interest) and *Commission* (misclassifying the pixel as the class of interest) instead. Their correspondences are: $Omission = 1 - Recall$ and $Commission = 1 - Precision$. We use *Precision* and *Recall* because they more directly convey the ability of the product to mask scenes. A number closer to 1.0 corresponds to better performance.

1.2.3.1 Using Landsat 8 Data for MOD09GA Error Assessment

To create a validation dataset directly with MODIS data, the 250 to 1000 m spatial resolutions of MODIS land bands (Table 1) are too coarse for consistent, accurate manual validation. The boundaries of clouds and snowpacks are hard to identify in the same contextual way a human analyst can with the spatial and contextual information in a Landsat scene with 30 m spatial resolution. The similar overpass times of MODIS Terra and Landsat 8 enable the use of Landsat 8 validation data to assess the MOD09GA cloud mask. While the clouds are not exactly in the same places in the two acquisitions, the large MODIS pixel size buffers against cloud movement, and the typical 20-minute gap between acquisitions means most clouds are nearly in the same locations. Only MODIS pixels that are fully covered by pixels from the Landsat reference mask are used. After alignment and removal of partially covered MODIS pixels, we were left with 21,023 manually classified MODIS pixels. The Landsat reference mask is ternary, with each pixel either *cloud*, *snow*, or *neither*. In coarsening the Landsat mask to MODIS pixel size, the resulting pixels have fractional covers of each category. We therefore classified the MODIS pixel as cloud or snow-covered if it had fractional coverage of 50% or more of Landsat reference mask pixels of that class. This coarsening resulted in 2,839 *snow*, 1,881 *cloud*, and 16,249 *neither* MODIS reference mask pixels. The MOD09GA cloud mask flags pixels as *clear, cloudy, mixed,* or *not sure.* We combined the MOD09GA classifications: pixels flagged as *cloudy* or *mixed* as cloud-covered and those flagged as *not sure* or *clear* as clear. The "not sure" pixels comprised 3.6% of the dataset.

1.2.3.2 Analyses of Reflectance Spectra

A plausible explanation for the difficulty that algorithms have in masking clouds over snow-covered terrain is that some clouds and some snow are spectrally too alike to discriminate, specifically for algorithms that make classification decisions from spectral thresholding and index tests. In our analysis, we highlight the difficulty in separating clouds and snow by extracting all snow and cloud spectra from the Landsat 8 validation data and exploring the similarities and differences between the snow and cloud classes.

To create these snow and cloud spectral libraries for analysis, we used the dataset of manually masked Landsat scenes to extract top-of-atmosphere reflectance spectra of the snow and cloud pixels from all the scenes. This extraction resulted in spectral libraries of 4.6 million snow pixels and 3.3 million cloud pixels.

To test the hypothesis that most snow and cloud pixels are spectrally distinct, but some are similar, we performed k-means unsupervised classifications of equally sized random subsets of the snow and cloud pixels from the spectral library. K-means classification (Hastie et al., 2009) maximizes similarity within each cluster while maximizing dissimilarity between clusters. Using k-means in this manner gives an overview of similarity between known classes of pixels. By running the same classification on different random subsets of pixels, we highlight the similarity or dissimilarity in spectral signatures between the classes. Because equally sized subsamples of pixels from the two spectral libraries are used, two classes of pixels that are inseparable and spectrally identical would lead to two clusters output from the classifier that are a 50/50 mix of both class types. Conversely, two classes of pixels that

are spectrally unique will be output from the classifier as two clusters that each have 100% membership from one of the input classes. How various runs compare between these two extremes enables us to determine what may lead to the similarity or dissimilarity between classes.

Snow and cloud pixels were classified with k-means using the spectral angle as the distance metric (Kruse et al., 1993) and the number of output clusters set to two. Pixels are clustered based on their similarity in angle away from the origin in multidimensional space, where each Landsat 8 band reflectance is a dimension. Given two spectral vectors $\overrightarrow{\boldsymbol{R_1}}$ and $\overrightarrow{\boldsymbol{R_2}}$, the cosine of the spectral angle $\angle_\lambda$ is the dot product of the two vectors divided by the product of their Euclidean norms:

$$\angle_\lambda = \cos^{-1}\left(\frac{\overrightarrow{\boldsymbol{R_1}} \cdot \overrightarrow{\boldsymbol{R_2}}}{\left\|\overrightarrow{\boldsymbol{R_1}}\right\|_2 \times \left\|\overrightarrow{\boldsymbol{R_2}}\right\|_2}\right) \quad (4)$$

Two pixels of the same class with reflectances of different magnitudes but similar spectral shape have similar spectral angles.

The magnitudes of the reflectances may be as important as the spectral angle for the most similar snow and cloud pixels. When detecting clouds or snow in Landsat 8 OLI bands 6 and 7, ice or water absorption dominates the signal with similar spectral shape, yet reflectance magnitude is sensitive to the size of the droplet or crystal. To test if the absolute magnitude of the signal is also important for snow and clouds, a second run of k-means was performed using the Euclidean norm of the differences, $\left\|\overrightarrow{\boldsymbol{R_1}} - \overrightarrow{\boldsymbol{R_2}}\right\|_2$, as the metric for clustering and the number of output clusters again set to two. Landsat OLI bands 1-7 were used for these classifications. Band 9 was omitted because the top-of-atmosphere cirrus band reflectance depends on the atmospheric

water vapor above each pixel; band 9 mis-identifies many high-elevation snow-covered pixels as cirrus.

We performed three runs of k-means classification: all snow and cloud pixels, snow and similar cloud, and the snow and dissimilar cloud. “Similar” cloud pixels have SWIR reflectances (bands 6 and 7) that are less than or equal to the 99th percentile of snow reflectances in the same bands from the snow pixel library. 35% of the cloud spectral library met these criteria and were identified as similar cloud and snow. We used the 65% of the cloud spectral library that were excluded from this dataset as the data for the unsupervised classification of dissimilar cloud and snow. Results from the unsupervised classification were compared with Precision, Recall, Overall Accuracy, and Separation, where the cluster is classed as the majority fraction of reference pixels it contains. Precision and Recall are the same tests used to assess the errors in the snow and cloud masks, Separation is the distance between the centroids of the two clusters using the same distance metric that the classification was run on. Because we have restricted the classification to two classes of the same size, Overall Accuracy $(TP + TN)/(TP + TN + FP + FN)$ is also a useful measure.

1.2.4 Modeling Reflectance of Snow and Cloud Pixels

We model the spectral signatures of water clouds, ice clouds, and mixed clouds based on the refractive indices of ice and water, the sizes of the water droplets and ice crystals, the clouds’ water equivalent, wetness in the case of mixed clouds, the illumination geometry, and the reflectance of the surface beneath the clouds. We model the spectral signature of snow based on effective optical radius, dust size and

concentration, the fractional cover of the snow in the pixel, and the reflectance of the surface comprising the remainder of the pixel. For each pixel analyzed, we invert the models to estimate cloud properties and snow properties that match the measured reflectances best. The physically based scattering model (Warren & Wiscombe, 1980; Wiscombe & Warren, 1980) is forced by the Mie scattering parameters for a sphere (Bohren & Huffman, 2007; Nussenzveig & Wiscombe, 1980) and the complex refractive indices of ice (Warren & Brandt, 2008) and water (Hale & Querry, 1973; Kou et al., 1993) to generate snow and cloud spectra with spherically equivalent radii that encompass the size distributions in clouds and snow. The Simple Model of the Atmospheric Radiative Transfer of Sunshine (SMARTS: Gueymard, 2005) generates the irradiance spectra illuminating the modeled snowpack along with atmospheric transmittance and thereby top-of-atmosphere reflectance. The snow spectra are linearly mixed with soil and vegetation spectra from the ECOSTRESS spectral library (Baldridge et al., 2009; Meerdink et al., 2019) to compute the reflectance of partially snow-covered pixels. The modeled clouds are not opaque, and the same spectral library is used to set the background reflectance underneath clouds.

We test the hypothesis that some snow and cloud pixels can have spectra from Landsat 8 OLI that match pixels of the opposite class. Snow pixels that CFMask classes as clouds and cloud pixels that CFMask classes as snow are considered in inversion models to solve for the best matches of cloud and snow properties that match the measurements. The inverted properties are then used to model the reflectances throughout solar spectrum at the wavelengths of the AVIRIS-NG airborne instrument (JPL AVIRIS Team, 2018), 380 to 2500 nm at 5-6 nm resolution, to investigate if

measurements from a spectrometer would distinguish snow from clouds, whose discrimination might be missed by Landsat 8 OLI.

1.3 Results

1.3.1 Assessment of Errors in CFMask and MOD09GA Cloud Mask

Figure 2 shows three examples of our approach for analyzing cloud masks. provides error statistics for all 26 scenes, along with each acquisition's terrain characterized by the percent cover of mountain landforms from the global Hammond landform regions (Karagulle et al., 2017), which quantifies Earth's mountainous locations at 250 m horizontal scale. Two of the USGS SPARCS masks with snow have no mountain cover, and their inclusion expands the range of snow spectra analyzed. One was an acquisition at high illumination angle from the Greenland ice sheet. The second is of cropland with variable snow-covered area. A USGS SPARCS mask of the Antarctic Peninsula Cordillera is outside the boundary of the Hammond dataset, so no mountain fraction was calculated. The scenes show differing proportions of cloud versus snow. The left column in Figure 2, from the New Zealand Alps, has roughly equal fractions of cloud and snow. The middle column from the Assam Himalaya in Bhutan has many clouds. The right column from the Bernese Alps in Switzerland consists mostly of snow. The images show the error results for clouds (TPcloud, TNcloud, FPcloud, FNcloud) with a class identifier for pixels that were not cloudy. *TNcloud_is_neither* pixels are clear-sky snow-free pixels and *TNcloud_is_snow* pixels are clear-sky snow-covered pixels, with neither class mapped as cloud by the satellite. *FPcloud_is_neither* pixels are clear-sky snow-free pixels and *FPcloud_is_snow* pixels are clear-sky snow-covered

pixels, with both classes mapped as cloud by the satellite. *FNcloud* are cloud pixels mapped as something else, *FNcloud_classed_snow* are the subset of this group that the satellite mapped as snow.

Table 5 shows the performance of the Landsat 8 CFMask and the MOD09GA cloud mask aggregated across all test scenes for the classes of interest, along with the ranges of values among all scenes. Generally, the *Precision* and *Recall* for the CFMask are high, whereas the *Recall* performance for the snow mask is lower. The MOD09GA mask does not include a snow class. In comparison to the CFMask, the MOD09GA cloud mask has a high *Recall* value (i.e., it sees most of the clouds) but low *Precision* (many false positives). The absolute value of each scene's central latitude was used to determine the proportion of variance in error statistics explained by scene location. The day of the water year, starting October 1st for northern hemisphere scenes and April 1st for southern hemisphere scenes, was used to determine the proportion of variance in error statistics explained by seasonality. R^2 values for both tests showed no significant attribute of error statistics explained by season or location.

1.3.2 Reflectance Spectra of Manually Classified Pixels

Figure 3 shows histograms for top-of-atmosphere reflectances in Landsat OLI bands 1-7 and 9 for the manually classified cloud and snow pixels, along with the histograms for the misclassified pixels. The figure indicates that most snow and cloud spectra are easily distinguishable. Snow is brighter in visible bands and darker in SWIR bands than most cloud spectra. Yet despite the general differences, spectral characteristics of numerous pixels in each class overlap.

1.3.2.1 Cirrus Band Reflectance from Snow at High Elevation

The wavelengths in the cirrus bands, Landsat OLI band 9 and MODIS band 26 (Table 1), experience absorption by atmospheric water vapor, so reflectance values in these bands are intended to map clouds high in the atmosphere. However, snow at high elevation also lies above most of the atmospheric water vapor, so the cloud masks based on putative cirrus observations often identify snow as clouds. One test for "potential cloud pixel" in the CFMask algorithm treats a top-of-atmosphere reflectance of 0.01 as having a 25% probability of identifying cirrus and a reflectance of 0.04 or greater with a 100% probability of cirrus (Zhu et al., 2015). These probabilities accumulate with the cloud probability scores from temperature and spectral variability. Figure 4 shows Landsat 8 OLI imagery of a 37×40 km area in Himachal Pradesh, India, acquired on 8 February 2016 in support of the ISRO-NASA AVIRIS-NG Indian campaign (Space Applications Centre, 2017). Elevations in the area range from 3100 to 6100m, and nearly all band 9 top-of-atmosphere reflectances exceed the CFMask's 0.04 threshold. Reflectance values from MODIS band 26 would likely be similar.

1.3.2.2 Similar Spectra of Snow and Clouds

Top-of-atmosphere reflectances in Landsat OLI bands 6 and 7 for dark clouds are below the 99th percentile of the snow spectra in these two bands (this reflectance value is about 0.2 for both bands). 35% of the clouds in the spectral library meet this criterion. Figure 5 shows examples of spectra modeled as snow and cloud within the Landsat 8 OLI bands and for the wavelengths of the AVIRIS-NG imaging spectrometer,

along with the region of interest in the Landsat image where the data were acquired. We identified the snow and cloud properties by solving for values for each that minimized the Euclidean norm of the differences between measured and modeled reflectances in Landsat OLI bands 1-7. From a random sample of ~1000 pixels with classification errors over the 26 Landsat scenes, we found that the Landsat spectral signature of 24% of the *FNcloud_classed_snow* pixels and 32% of the *FPcloud_snow* pixels matched plausible snow and cloud properties. All simulations considered a dark loam (Class Inceptisol, Subclass Xerumbrept) from the ECOSTRESS Library (Meerdink et al., 2019) as the surface underlying the cloud or occupying the snow-free part of the pixel).

Although the snow and cloud spectra are distinctive if the full spectrum is considered, in the wavelength bands of the OLI they are not, and the reflectance in the cirrus band (band 9) would not clarify the spectra as snow or cloud. An imaging spectrometer, as recommended in the recent Decadal Survey for Earth Science and Applications (National Academies, 2018), could distinguish between the snow and the cloud.

Traditional snow mapping methods often employ the Normalized Difference Snow Index (Dozier, 1989); typically the green and shortwave infrared bands are used, thus bands 3 and 6 for the Landsat 8 OLI, so $NDSI = (R_3 - R_6)/(R_3 + R_6)$. Figure 6 shows the range of NDSI values for various sets of pixels in the study. Snow with similar spectral characteristics to clouds has similar NDSI values, shown in the boxplots of dark cloud and false positive snow pixels, i.e., the snow pixels misclassified as cloud by CFMask. Clouds with similar spectral characteristics to snow have similarly

high NDSI values, shown in the similarity of boxplots of snow NDSI and the NDSI of cloud pixels that CFMask misclassifies as snow. Many of the false positive snow spectra are outliers within the context of the entire snow spectral library. Most snow has a high NDSI value between 0.5 and 1, distinguishing it from clouds, which tend to have NDSI values from –0.4 to +0.4. However, the outlier conditions that are confused with clouds comprise a source of error in scenes where the snow cover is composed of these more difficult-to-discriminate snow pixels.

Snow that covers an entire pixel is spectrally distinguishable from most clouds, but snow pixels that are fractional mixtures with other surface types can have spectra that are like some clouds. For fractional snow pixels, visible reflectance decreases, and SWIR reflectance increases as soil fraction increases. At the coarse spectral resolution of Landsat and MODIS, the spectra of the mixed "bright SWIR" snow pixels mimic some cloud spectra. The *FPcloud_snow* pixels at the borders of clouds in Figure 2 and the spectra in Figure 5 highlight this effect.

1.3.3 K-means Analyses

Table 3 shows k-means performance for three runs using spectral angle to measure the similarity of snow and cloud pixels and the single run using the Euclidean norm to measure the similarity among the 35% of cloud pixels that are "dark." The other 65% of cloud pixels are grouped as "bright clouds." The ability to separate snow from clouds decreases as the spectra become similar. *Separation*, *Precision,* and *Recall* decrease when the snow is classified with the similar "dark cloud" pixels. For these most similar snow and clouds, *Precision* of snow and *Recall* of clouds improve by using

a magnitude-based measure over a spectral angle-based measure, but performance is still worse than the other two spectral angle runs as these two sets of pixels, dark clouds and snow, are spectrally similar.

1.4 Discussion

1.4.1 Cloud Mask Performance

Cloud masking algorithms perform worse over snow-covered areas compared to other land surfaces because the multispectral sensor measurements can contain snow spectra indistinguishable from clouds and cloud spectra indistinguishable from snow. Significant improvement in our abilities to distinguish clouds from snow must utilize additional information beyond per pixel spectral characteristics.

Table 4 displays the results from prior cloud masking studies for comparison to the MOD09GA and CFMask performance measures from this study. CFMask performance is lower over snow-covered mountain regions compared to previous global assessments of Landsat 8 cloud masking. The *Recall* and *Precision* of CFMask over mountains is lower compared to prior error assessments in polar regions of CFMask, the best cloud mask in these regions tested by Foga et al. (2017). The MOD09GA cloud mask has lower performance metrics compared to the CFMask results of the same scenes and to all prior studies that assessed Landsat 8 cloud mask performance. Both MODIS and Landsat scenes exhibit a spread in performance based on the local composition of each scene. Therefore, depending on the spectra of the cloud and the snow in the scene, performance can vary from excellent to unacceptable.

The performance measures of CFMask to classify clear-sky pixels without snow exceeded 95%. Low fractional snow cover leads to "bright SWIR" snow pixels and can be difficult to detect by a human analyst, and the false positive cloud mask pixels from CFMask that were clustered around the borders of the snow suggests that many of these pixels contained fractional snow missed by the analysts. Users should be comfortable using CFMask for identifying clear-sky snow-free pixels in mountain scenes unless the region of interest is the boundary around a snowpack.

The MODIS MOD09GA cloud mask has worse performance than CFMask with low *Precision* and moderate *Recall*. One plausible factor for worse performance may have been the manner in which the OLI reference mask, which we coarsened to MODIS resolution using Gaussian pyramid reduction (Burt & Adelson, 1983), was used to assign a single reference class to each MODIS pixels. We tested OLI reference mask coverage thresholds of 10% and 100% for classing MODIS pixels. This variability had moderate impact on precision (0.223 and 0.0917) and minimal impact on recall (0.699 and 0.72).

The MOD09GA algorithm misclassifies many snow pixels in each scene as cloud, and many cloud pixels are missed and can be mistaken for snow. This pervasive misclassification presents a challenge for extracting snow data from MODIS imagery. The misclassification of snow as cloud and vice versa adds bias to models of snowmelt (Andreadis & Lettenmaier, 2006) and estimates of snow water equivalent (Rittger et al., 2016) that assimilate data from MODIS. Improvement of cloud masks over snow could improve these models.

One reason for the decrease in performance compared to the Landsat 8 product is that the 500-m cloud mask must classify more mixed pixels than the Landsat 8 30-m cloud mask. The spatial scale of MODIS pixels creates spectral mixture problems for more pixels. 60% of MODIS pixels for which the Landsat CFMask correctly classified cloud and snow had snow or cloud coverage fractions less than 100% coverage. The false positive rate of the MOD09GA cloud mask is also significantly higher than that of CFMask. Many clear-sky, bare ground pixels were included in the MOD09GA cloud mask, so while the remaining pixels may be bare, many useable pixels of surface reflectance are mistakenly masked as cloud, a problem that was not observed in Landsat 8 imagery. Users cannot assume that MODIS pixels are either entirely cloud-free or entirely cloud-covered in imagery over mountains. The mixed-pixel problem for cloud classification at 500-m resolution is fundamentally different than at 30 m for Landsat, as mixed cloud pixels are rare at the smaller Landsat spatial scale. These larger mixed pixels in MODIS reduce the sharpness of the spectral differences between pure endmembers of snow, clouds, and other surfaces.

We did not directly analyze snow masking abilities of the MOD9GA cloud mask, because the MODIS QA bands for snow detection do not confirm snow or ice but just flag the processing path of the cloud mask algorithm for that pixel. The MODIS Collection 6 snow mask is an ancillary data product unrelated to the cloud mask. The existing snow masking for MODIS relies on the cloud mask to identify cloud pixels, and the poor results of the MOD09GA cloud mask show that any analysis of snow that removes pixels that are flagged as clouds will likely elide valid snow-covered pixels.

The snow masking code in CFMask is based on NDSI and band 9 cirrus thresholds. Moreover, the mask is conservative towards clouds; if there is ambiguity in the algorithm logic, it classifies the pixel as cloud instead of snow. Many of the false positive snow pixels classed as cloud are outliers on the library of snow spectra, but many of these pixels are around the snowline. As the snow melts and disappears, the fractional snow-covered area of pixels decreases, and pixels transition from the interior of the snowpack to the boundary. These are important to class as snow for modeling snow water equivalent and snowmelt runoff.

1.4.2 Spectral Library Insights

This project benefited from an extensive set of reference data that included a snow class separate from other land surfaces to test cloud mask performance. The manually classified snow and cloud masks enabled us to construct a spectral library to explore if snow and cloud can have similar spectral signatures. Our spectral library is not necessarily a representative sample of the frequency and range of possible cloud and snow spectra encountered in remote sensing imagery. But, by looking at both scene-by-scene statistics and the library of all scenes' spectra together, our evaluations highlight the possibility for indiscriminate snow and cloud spectra within a scene.

The insight that spectral reflectance of 35% of the cloud pixels in the study resemble spectral reflectance of snow in the SWIR bands confirms that while most snow and cloud are spectrally distinct, the problem of spectral indiscrimination is not a fringe issue. The k-means analysis compares the similarity between the two classes based on a classification technique that relies exclusively on a calculation of spectral

similarity. The fractions of pixels within the spectral library that exhibit strong similarities to the general spectral characteristics of the opposite class is too large to be attributed to chance or random error. In situations where the reflectances in the SWIR bands are similar for snow and clouds, discrimination between them becomes difficult.

Spectra from pixels with large fractions of subpixel soil can resemble spectra of thin clouds, with the mixed pixels showing elevated SWIR reflectance compared to the snow within the mixture. Further anecdotal evidence lies in the band of misclassified clouds around the edges of the snowpack in some Landsat 8 images. The same pattern is observed in the MODIS data, where at 500 m many more of the snow pixels are mixtures, increasing the misclassification rate compared to Landsat. In addition to snow with elevated SWIR reflectance, clouds with low reflectance in SWIR bands are common and their NDSI values resemble those for snow. Reflectance in Landsat OLI band 9 or MODIS band 26 of snow at altitude will often have values like those of cirrus clouds.

1.4.3 Ideas for Improving Cloud Masks in Imagery from Multispectral Sensors

Resolving these types of discrimination issues requires new approaches to cloud masking that account for local viewing conditions, textural features, and persistence. This error analysis and spectral investigation helps formulate ideas for future improvements in cloud masking over snow-covered terrain. For snow-cloud discrimination in a single Landsat 8 or MODIS image, textural features might be used; our dataset provides information to develop such approaches. Mountainous terrain

itself provides spatial patterns in imagery, so one potential approach is to see whether those patterns cannot be recognized in an image because of obscuration by clouds.

The twice daily MODIS daytime acquisitions allow cloud discrimination by examining the time series, applying various filters, and interpolating to fill the gaps caused by obscuring clouds (Gafurov & Bárdossy, 2009; Helfrich et al., 2007; Tran et al., 2019). Sentinel 2a&b along with Landsat 8 provide coverage at repeat intervals of 4 days. Their coincident overpass times with MODIS Terra, could offer additional validation data for MODIS and perhaps a source of training data for machine learning approaches to cloud masking with MODIS. For snow in forests, Rittger et al. (2019) identify clouds by detecting changes in the estimated fraction of the vegetation or soil endmember in pixels with fractional snow. Detection of cloud movement would be especially useful with data from geostationary satellites. The improved capabilities of the generation of GOES-16/17, Himawari-8/9, and FengYun-4 satellites show promise for leveraging the spectral capabilities of MODIS and its successors with the high temporal frequency of geostationary satellites to mask clouds over snow.

Going forward, a reliable cloud mask is necessary to create the best snow mask. By including currently missed clouds in a cloud mask and excluding the misclassified snow, existing snow mapping techniques that are reliably used under clear-sky conditions could be extended with similar performance to cloudy scenes.

1.4.4 The Case for Imaging Spectroscopy

Among the five observations in the "designated" (high priority) class recommended by the National Academies of Science, Engineering, and Medicine for

future space missions is imaging spectrometry in the visible and shortwave infrared and multispectral imagery in the thermal infrared (National Academies, 2018). As Figure 4 through Figure 6 show, existing multispectral sensors sometimes have similar spectral signatures for snow and clouds. Measurements of spectral reflectance across the solar spectrum from 400 to 2500 nm at 5-10 nm resolution would enable unambiguous discrimination with reliable estimates of soil or vegetative background reflectance, the sizes of the scattering ice crystals or water droplets, fractional snow cover, thickness of clouds, and degradation of snow albedo by dust or soot.

1.5 Conclusion

Satellite measurements provide our best resource for furthering our understanding of the cryosphere. Cloud masking algorithms over snow-covered terrain perform poorly compared to over other land surfaces, and existing cloud masking performance depends on local cloud and snow types and cannot deliver consistently accurate results. In multispectral images of the visible and infrared spectra, many clouds and snow pixels are unambiguously separable, mainly for thick clouds and snow that fully covers a pixel. However, variations in snow and cloud properties can cause overlap in their spectral signatures within a scene. While multispectral differences alone can identify many pixels correctly as cloud or snow, others remain ambiguous. Landsat 8 OLI data confirm the possibility of cloud spectra like snow or snow spectra like cloud as a plausible source of algorithm misclassifications. Availability of imagery from a future spaceborne imaging spectrometer would provide validation data to evaluate new multispectral methods for snow-cloud discrimination.

1.6 Data

This chapter does not exhaust the ways in which the data we have assembled could be used to understand snow-cloud discrimination. Although I explore new methods using these data in chapter 2, all data that we created are available (at Stillinger & Collar, 2019) to other researchers interested in the same problem. Table 5 lists all 26 Landsat 8 OLI and corresponding MODIS images analyzed in this study. The data we provide with unrestricted access include the geoTIFFs where we have manually identified snow and clouds in 13 images. The collection from the U.S. Geological Survey (2016) includes full-resolution images from their manual classification. We also provide the spectral libraries for clouds, snow, false positives, and false negatives in the manually classified imagery.

MATLAB code for the snow albedo model (based on Wiscombe & Warren, 1980) is available on GitHub at https://github.com/edwardbair/SCAGD/. The SMARTS model for atmospheric radiation is available from the National Renewable Energy Laboratory: https://www.nrel.gov/grid/solar-resource/smarts.html. The ECOSTRESS spectral library of natural and human-made materials is available from JPL, https://speclib.jpl.nasa.gov/.

1.7 Tables

Table 1: Spectral intervals (full-width half-maximum) and spatial resolution for the Landsat 8 OLI (Barsi et al., 2014; NASA/USGS Landsat 8 Team, 2014) and the MODIS "land" bands (NASA MODIS Team, 2003).

	band	vavelength range, nn		spatial resolution, m	description
Landsat OLI	1	435	451	30	ultra blue
	2	452	512	30	blue
	3	533	590	30	green
	4	636	673	30	red
	5	851	879	30	near-infrared
	6	1567	1651	30	shortwave-infrare
	7	2107	2294	30	shortwave-infrare
	8	503	676	15	panchromatic
	9	1363	1384	30	cirrus
MODIS "land" bands	1	620	670	250	red
	2	841	876	250	near-infrared
	3	459	479	500	blue
	4	545	565	500	green
	5	1230	1250	500	near-infrared
	6	1628	1652	500	shortwave-
	7	2105	2155	500	shortwave-
	26	1360	1390	1000	cirrus

Table 2: Top panel shows summary statistics for CFMask and MOD09GA cloud mask performance in the test scenes for each class, by aggregate of all scenes and median, minimum, and maximum values for individual scenes. Bottom panel shows the R^2 values for individual scene errors compared to the scenes' latitudes and seasonality

	Landsat 8 CFMask				**MOD09GA Cloud Mask**			
Aggregated scenes					Aggregated scenes			
	Precision	*Recall*	*F statistic*			*Precision*	*Recall*	*F statistic*
Cloud	0.695	0.857	0.768		Cloud	0.166	0.723	0.270
Snow	0.938	0.787	0.856					
Neither	0.943	0.975	0.959		Neither	0.964	0.602	0.741
Scene-by-scene median (min, max)					Scene-by-scene median (min, max)			
	Precision	*Recall*	*F statistic*			*Precision*	*Recall*	*F statistic*
Cloud	0.803 (0.03, 0.99)	0.793 (0.03, 0.98)	0.771 (0.03, 0.96)		Cloud	0.111 (0.00, 1.00)	0.833 (0.00, 1.00)	0.156 (0.00, 1.00)
Snow	0.897 (0.65, 1.00)	0.449 (0.00, 1.00)	0.536 (0.00, 0.98)					
Neither	0.850 (0.00, 0.99)	0.929 (0.28, 1.00)	0.867 (0.00, 0.99)		Neither	0.977 (0.18, 1.00)	0.415 (0.00, 1.00)	0.495 (0.00, 1.00)

	Latitude vs Error (R^2)		**Season vs Error (R^2)**	
	Landsat 8 OLI	*MOD09GA*	*Landsat 8 OLI*	*MOD09GA*
Precision	0.05	0.10	0.01	0.00
Recall	0.03	0.01	0.22	0.01
F statistic	0.01	0.01	0.03	0.00

Table 3: Performance statistics for the unsupervised classification of the snow and "dark clouds," snow and all clouds, snow and "bright clouds." The separation distance metric for the Euclidean norm is excluded as no comparative runs were needed.

				Precision		Recall	
Method	Spectral Libraries	Separation	Overall Accuracy	Snow	Cloud	Snow	Cloud
Spectral Angle	Snow / Bright Cloud	0.09	0.957	0.961	0.953	0.953	0.961
	Snow / Cloud	0.08	0.918	0.894	0.945	0.948	0.887
	Snow / Dark Cloud	0.07	0.850	0.796	0.927	0.940	0.759
Euclidean Norm	Snow / Dark Cloud		0.788	0.841	0.744	0.698	0.878

Table 4: Comparison of error statistics for CFMask and the MOD09GA cloud mask to published error statistics.

	Precision	Recall
MOD09GA cloud mask in snow-covered mountains (this stud	0.166	0.723
CFMask for clouds in snow-covered mountains (this study)	0.695	0.857
CFMask for snow and ice (Foga et al., 2017, Table 7)	0.757	0.866
CFMask global assessment (Foga et al., 2017, Table 5)	0.880	0.973
Fmask global sample (Hughes & Hayes, 2014)	0.740	0.986
ACCA "best cloud mask over snow/ice" (Foga et al., 2017, Tal	0.706	0.993

Table 5: Statistics for Precision, Recall, and the F-statistic for CFMask for all Landsat 8 OLI images and the MODIS cloud mask for all MOD09GA images analyzed

Date	Sensor/ Product	Scene Coordinates	Mountain Cover	Geographic Coordinates		*Precision*	*Recall*	*F statistic*
13-Jan-2014	Landsat 8 OLI	WRS P 76 R 91	67%	-45	168.3	0.765	0.892	0.824
	MOD09GA	MODIS h29v13 & h30v13				0.246	0.845	0.381
15-Apr-2014	Landsat 8 OLI	WRS P 137 R 41	90%	27.8	90.8	0.856	0.965	0.907
	MOD09GA	MODIS h26v06				0.553	0.997	0.711
5-Jun-2013	Landsat 8 OLI	WRS P 195 R 28	89%	46.4	7.8	0.766	0.939	0.844
	MOD09GA	MODIS h18v04				0.101	0.907	0.182
1-Apr-2014	Landsat 8 OLI	WRS P 167 R 51	60%	47.8	-113.9	0.992	0.801	0.887
	MOD09GA	MODIS h22v07					0.000	0.000
1-May-2014	Landsat 8 OLI	WRS P 153 R 37	79%	49.0	-113.6	0.934	0.896	0.915
	MOD09GA	MODIS h23v05				0.018	1.000	0.036
4-Dec-2014	Landsat 8 OLI	WRS P 152 R 35	100%	39.3	-107.7	0.999	0.726	0.841
	MOD09GA	MODIS h23v05				0.084	1.000	0.156
4-Jun-2013	Landsat 8 OLI	WRS P 35 R 33	95%	51.8	-72.9	0.960	0.810	0.879
	MOD09GA	MODIS h09v05				0.003	0.100	0.005
4-Jun-2013	Landsat 8 OLI	WRS P 35 R 34	76%	49.7	-122.2	0.969	0.861	0.912
	MOD09GA	MODIS h09v05				0.000	0.000	0.000
5-Apr-2014	Landsat 8 OLI	WRS P 42 R 34	89%	36.7	76.3	0.998	0.905	0.949
	MOD09GA	MODIS h08v05				0.000	0.000	0.000
6-Sep-2014	Landsat 8 OLI	WRS P 217 R 15	35%	27.7	84.8	0.034	0.030	0.032
	MOD09GA	MODIS h17v02				0.228	0.676	0.341
7-May-2014	Landsat 8 OLI	WRS P 42 R 33	87%	27.8	90.8	0.747	0.906	0.819
	MOD09GA	MODIS h08v05				0.227	0.833	0.357
7-May-2014	Landsat 8 OLI	WRS P 42 R 34	89%	36.7	76.3	0.802	0.859	0.829
	MOD09GA	MODIS h08v05				0.202	1.000	0.337
10-Dec-2013	Landsat 8 OLI	WRS P 231 R 94	80%	30.4	78.2	0.778	0.894	0.832
	MOD09GA	MODIS h13v13				0.165	0.847	0.276
11-Apr-2013	Landsat 8 OLI	WRS P 146 R 39	75%	46.5	7.8	0.922	0.896	0.909
	MOD09GA	MODIS h24v05 & h24v06				0.833	1.000	0.909
14-Aug-2014	Landsat 8 OLI	WRS P 47 R 25	98%	33.8	74.3	0.792	0.953	0.865
	MOD09GA	MODIS h10v04				0.564	0.872	0.685
14-Jul-2013	Landsat 8 OLI	WRS P 148 R 35	100%	64.8	-17.0	0.865	0.780	0.820
	MOD09GA	MODIS h24v05				0.111	0.038	0.057
14-May-2013	Landsat 8 OLI	WRS P 153 R 35	94%	37.5	-108.6	0.984	0.595	0.742
	MOD09GA	MODIS h23v05					0.000	0.000
21-Jun-2014	Landsat 8 OLI	WRS P 5 R 15	0%	-70.4	-65.8	0.707	0.136	0.228
	MOD09GA	MODIS h15v02				0.013	0.059	0.021
21-Mar-2014	Landsat 8 OLI	p170r025	0%	38.4	-119.4	0.685	0.670	0.677
	MOD09GA	MODIS h20v03 & h21v03				0.000	0.000	0.000
26-Dec-2013	Landsat 8 OLI	WRS P 215 R 110	-*	37.6	-118.3	0.056	0.860	0.106
	MOD09GA	MODIS h15v16						
26-May-2014	Landsat 8 OLI	WRS P 15 R 24	8%	63.7	-49.3	0.706	0.906	0.794
	MOD09GA	MODIS h13v03				0.240	0.826	0.372
28-Dec-2013	Landsat 8 OLI	WRS P 149 R 37	67%	-49.1	-73.0	0.807	0.852	0.829
	MOD09GA	MODIS h24v05				0.317	0.626	0.421
29-May-2013	Landsat 8 OLI	WRS P 41 R 27	96%	-44.5	168.3	0.991	0.866	0.925
	MOD09GA	MODIS h10v04				0.066	1.000	0.123
29-May-2013	Landsat 8 OLI	WRS P 146 R 39	75%	46.5	7.8	0.883	0.894	0.888
	MOD09GA	MODIS h24v05 & h24v06				1.000	1.000	1.000
30-Apr-2015	Landsat 8 OLI	WRS P 141 R 41	65%	50.5	47.5	0.940	0.984	0.961
	MOD09GA	MODIS h25v06				0.018	1.000	0.036
30-Jun-2013	Landsat 8 OLI	P 41 R 26	59%	-44.5	168.3	0.940	0.737	0.826
	MOD09GA	MODIS h10v04				0.000	0.000	0.000

1.8 Figures

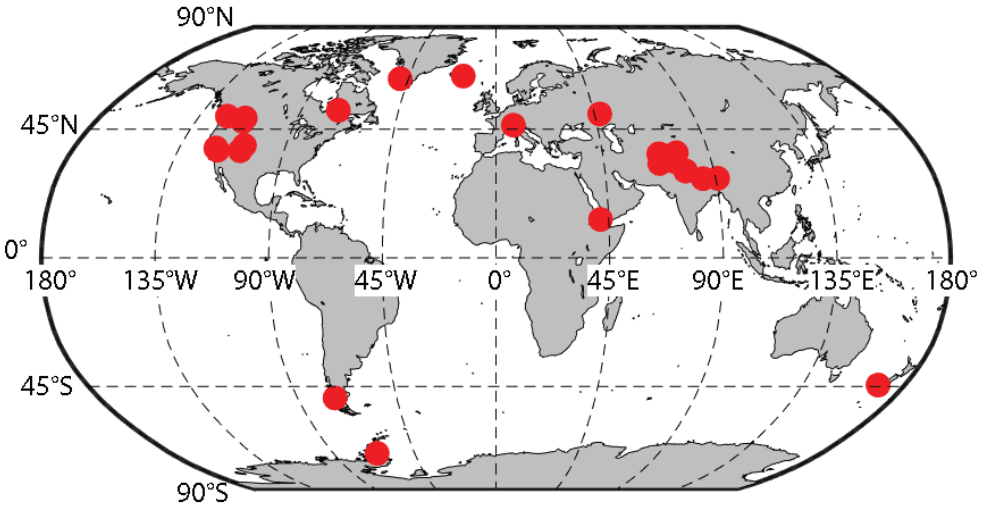

Figure 1: Red dots show the locations of the Landsat 8 validation images.

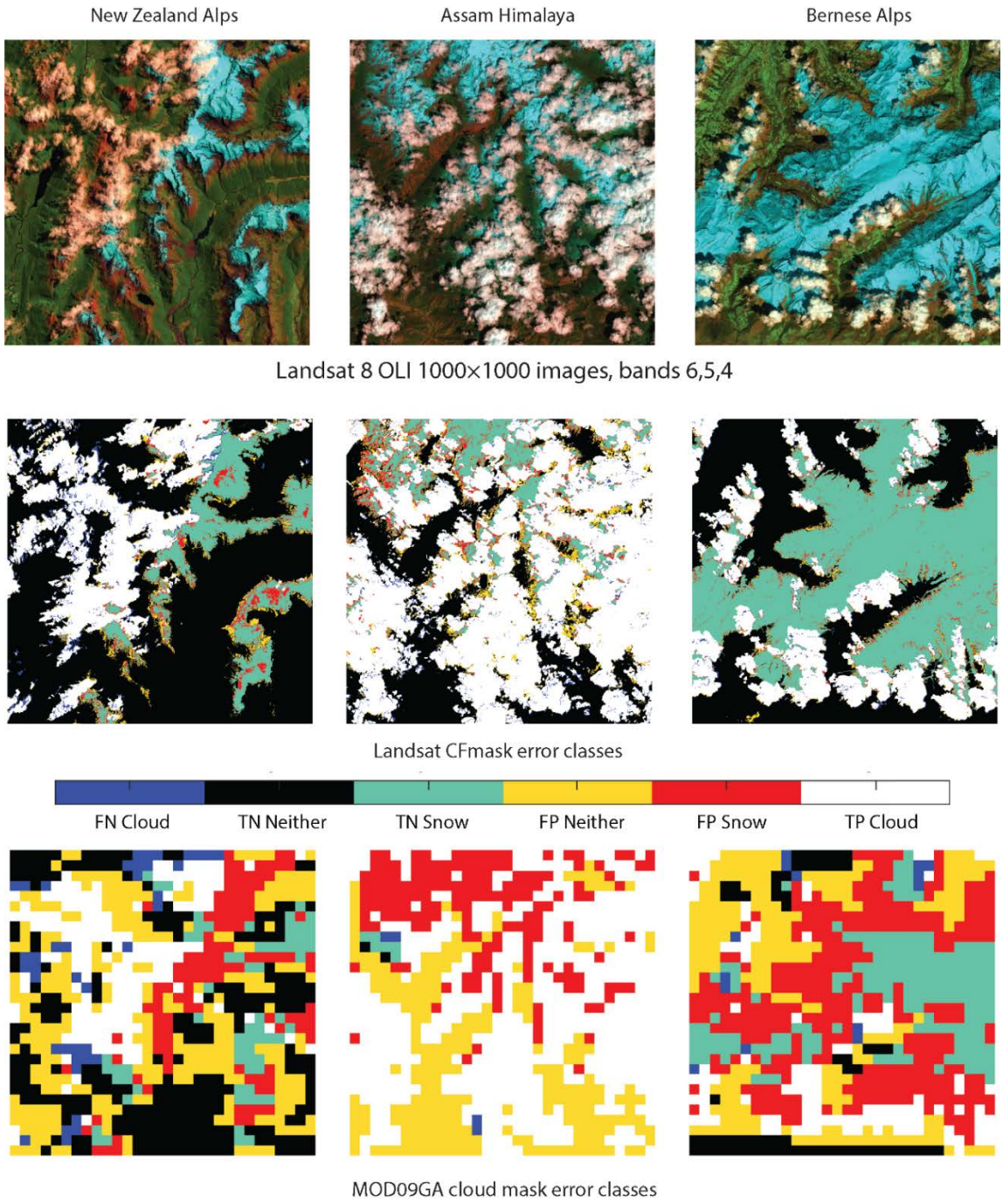

Figure 2: Examples of CFMask and MOD09GA cloud mask performance over three areas with varying degrees of cloud cover. Pixels are colored by surface type and identified as false negative (FN), true negative (TN), false positive (FP) or true positive (TP) in the c cloud mask. Left column New Zealand Alps: Climax Peak (2446 m) in the Pyke and Dart River Basins, clouds and snow about equal. Middle column Assam Himalaya: below Chura Kang (6300 m) in Bhutan, cloud cover extensive but snow visible. Right column Bernese Alps: Lötschental and Bietschhorn (3924 m), mostly snow with few clouds.

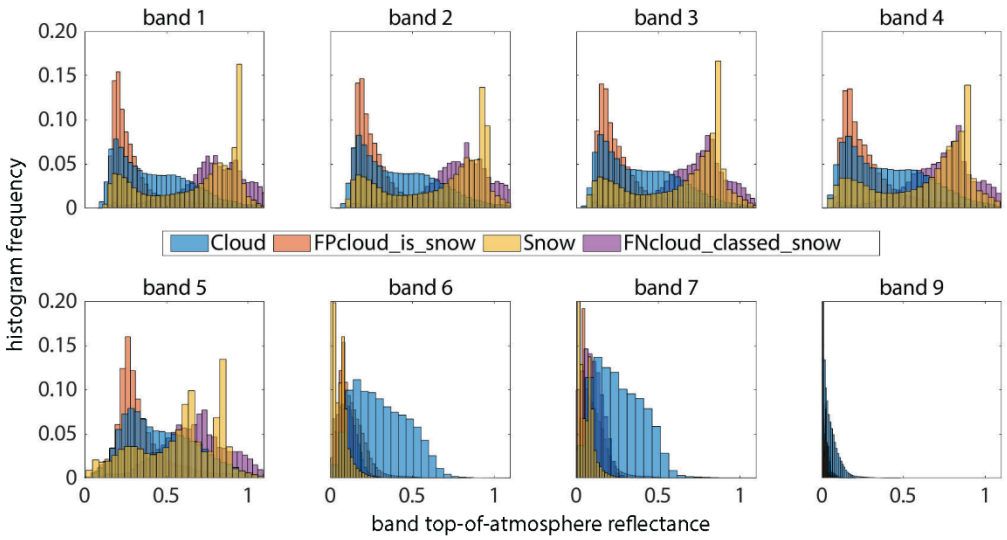

Figure 3: Normalized histograms of the top-of-atmosphere reflectance in Landsat 8 OLI bands 1-7 and 9 for manually classified snow and cloud pixels, along with histograms for misclassified pixels, combining snow classified as cloud (FPcloud_is_snow) and cloud classified as snow (*FNcloud_classed_snow*). The snow spectral values show a bimodal distribution with one mode in line with the skewed distribution of the cloud spectra.

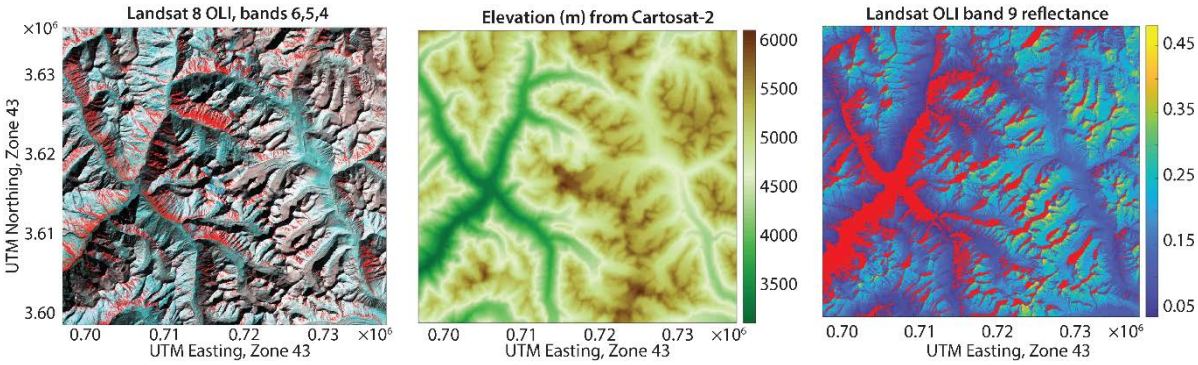

Figure 4: Left image shows false-color rendition of Landsat 8 OLI image on 8 February 2016 in the Himachal Pradesh, India. Middle image shows corresponding elevations from Cartosat-2. Right image shows values of Landsat 8 OLI band 9 top-of-atmosphere reflectance that exceed 0.04, with values less than 0.04 masked in red. Values above 0.04 would be classified as cloud even though there are no clouds in the frame. The use of band 9 to identify clouds will identify snow at high elevation as clouds.

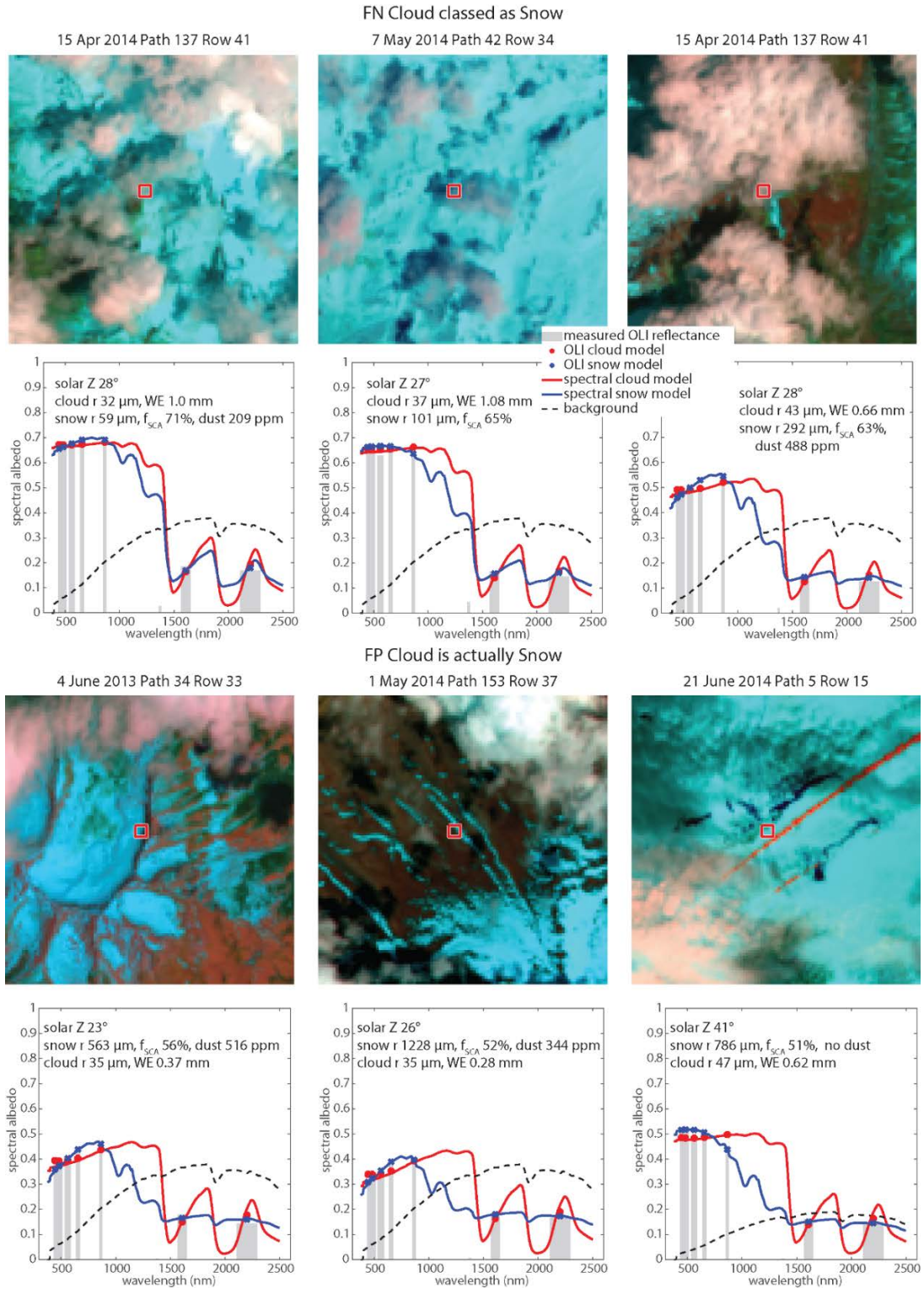

Figure 5: Examples of similar snow and cloud spectral signatures in the Landsat 8 OLI bands, chosen for illustration from many examples. The upper images show FNcloud_classed_snow (clouds were missed because they were identified as snow); the lower row of images shows *FPcloud_is_snow*

(snow was misclassified as cloud). In the corresponding graphs, the grey bars show measured OLI spectra. The symbols show these spectra modeled as snow (blue) and ice clouds (red) with minuscule (10^{-8}) liquid water concentration, and the text within the figures identifies the snow and cloud properties that match the spectra. f_{SCA} is the fractional snow cover of the pixel, and WE is the cloud water equivalent. The background reflectance is for a dark loam (Class Inceptisol, Subclass Xerumbrept). The reflectances in the Landsat bands are similar enough to obviate snow-cloud discrimination. The blue and red lines show they would be distinguishable with continuous measurements over the wavelengths of the AVIRIS-NG sensor, 380 to 2500 nm at 5-6 nm resolution.

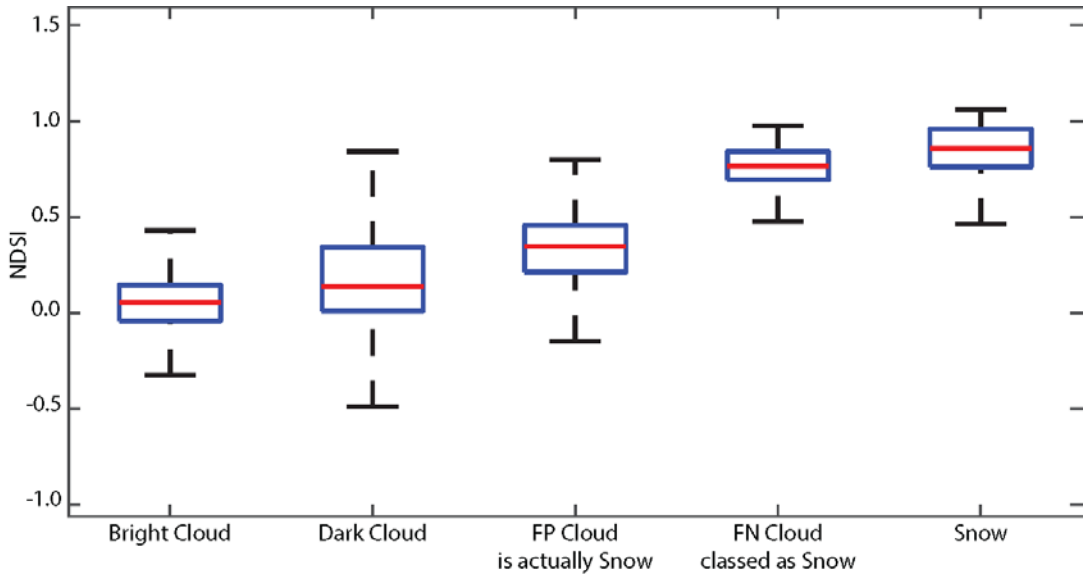

Figure 6: NDSI calculations for subsets of the snow and cloud spectral libraries. Outliers are not plotted, but the box-and-whisker plot shows overlaps among the classes. Bright clouds and snow differ significantly, whereas dark clouds and snow show enough whereas dark clouds and snow show enough overlap to cause misclassification. Cloud pixels classified as snow by CFMask and true snow pixels are spectrally similar. Clouds can look like snow and snow can look like clouds.

2 Multispectral Cloud Mask with Optical Properties

2.1 Introduction

To observe Earth, satellites from countries around the world image the planet each day. In these images, clouds are a main feature of Earth's atmosphere; globally covering around 70% of Earth's surface (Stubenrauch et al., 2013). For observing both the clouds and the surfaces below, accurate cloud masking is a necessary step in the image processing workflow.

Clouds are bright and white in visible spectra and cold in the thermal infrared. These characteristics distinguish them from most surfaces on Earth, yet snow measured from multispectral satellites can be spectrally indistinct from clouds (Stillinger et al., 2019). Because cloud and snow are composed of small particles of ice, water, or a mixture, the optical properties of water and ice explain their spectral signatures. For clouds with no impurities, the size and composition of the scattering particles and the cloud water equivalent determine the albedo and measured reflectance by remote sensors. The absorption coefficients k of water (Hale & Querry, 1973) and ice (Warren & Brandt, 2008) vary by eight orders of magnitude across the wavelengths in the solar spectrum. On average, both water and ice become more opaque with increasing wavelength, and in the near-infrared (0.75-1.1 μm) and shortwave-infrared (1.1-2.5 μm) parts of the spectrum, the sizes of the scattering water and ice particles are similar to the scaling length of the absorption, defined as $\lambda/(4\pi k)$, where λ is the wavelength of the radiation. Size differences between cloud and snow scatterers in these wavelengths are the main feature that enables snow-

cloud discrimination (Dozier, 1989), whereby the larger snow grains reflect less radiation than the smaller cloud particles. While most clouds are brighter in shortwave-infrared wavelengths than snow, some mixed pixels of snow and soil or vegetation can have multispectral signatures that are similar to those for some thin clouds (Stillinger et al., 2019). These spectrally similarities of pixels limit the performance of cloud masks that class pixels individually.

Most traditional cloud masking methods apply spectral tests to individual pixels to identify clouds. The original Landsat Automated Cloud Cover Assessment (ACCA) algorithm (Irish et al., 2006) provided users an estimate of cloud cover of a scene. It helped users select scenes by flagging images that might be too contaminated by clouds. Although ACCA can reliably identify clouds in scenes without snow or ice, this algorithm has difficulties in snow- and ice-covered regions and in accurately masking the precise cloud objects in a scene. With the wide accessibility of Landsat data, accurate cloud masks are necessary (Zhu & Woodcock, 2014b). Especially in the mountains, most scenes contain many clouds, hence there is a need to identify cloud-free pixels even in scenes with extensive cloud cover.

In 2012, Fmask became the new standard for cloud mask products in Landsat imagery. Over the years it has been improved and adapted to work on more sensors, including Sentinel-2a/b, and in mountainous terrain (Qiu et al., 2017; Zhu & Woodcock, 2012, 2014a; Zhu et al., 2015). In its current operational form, as CFMask, it is the cloud mask that is shipped in the quality file with Landsat data. The algorithm relies on spectral tests, the statistical distribution of reflectance across various surface types, and cloud-shadow matching. For areas without snow and ice, the algorithm performs

as well as human analysts. Its original intent is to identify pixels not contaminated by snow or clouds. From this perspective, CFmask shows the excellent abilities of a well-designed, spectrally based algorithm to identify clear sky pixels in multispectral sensors. Unfortunately for the cryosphere community, the CFMask algorithm suffers from errors of commission and omission between snow and clouds, and the product that ships with Landsat 8 data in the BQA file cannot reliably mask clouds over snow at the same level of accuracy as it can over other land covers (Foga et al., 2017; Stillinger et al., 2019).

2.2 Recent Approaches to Cloud Masking using Machine Learning

New cloud masks for Landsat 8 have come out in the last few years that utilize computer vision image segmentation techniques. The new methods are possible with the combination of easy access to Landsat imagery and a large set of accessible validation data from the Landsat 8 Cloud Cover Assessment (CCA) and USGS Spatial Procedures for Automated Removal of Cloud and Shadow (SPARCS) validation datasets (Foga et al., 2017; U.S. Geological Survey, 2016). The first widespread machine learning based algorithm was the SPARCS method (Hughes & Hayes, 2014). By using a validation data set of manually classified Landsat 8 images, they were able to train a neural network to classify snow and cloud pixels.

Li et al. (2017) used a combination of spectral tests and the notion that some clouds thin towards their edges to guide filtering of the clouds alongside differences in the shape and texture of bright features to exclude them from the cloud mask. They used spatial pattern analysis and local binary patterns to extract geometric and

textural information and improve the cloud mask. In a subsequent cloud mask, Li et al. (2019) used deep learning methods to train a cloud classifier that performs well on a range of multispectral sensors using an encoder-decoder architecture and 516 training images.

An alternative approach to starting with individual pixels is to use superpixels (segmented regions of local similarity) as the basic size of the data for which feature sets are created. Zi et al. (2018) created superpixels, then classified these into non-cloud, thin cloud, and thick cloud via a convolutional neural network (CNN) where the feature set for the CNN are characteristics for each superpixel. After classifying superpixels via CNN, the superpixel boundaries are refined with conditional random fields. Another recent approach (Chai et al., 2019) uses CCNs trained with the Landsat 8 CCA validation dataset. Landsat images are batched into smaller images and run through the CNN to classify them. The manually classified images are used in a 60/10/30 split for training, validation, and testing.

These latest Landsat cloud masking algorithms using CNNs to identify clouds in images have yielded impressive classification results of Landsat imagery but less progress in extracting additional meaningful information from the images. The physical meaning of the spectral characteristics of clouds and other land surfaces are bypassed by the neural networks. Once these algorithms are trained, they use semantic segmentation techniques that were originally designed for real time segmentation, so they run quickly. Faster can be better, but it is difficult for users to derive physical meaning from errors and results. Results depend on the training dataset, and any errors in manual classification by human analysts in the training datasets translate

directly into cloud mask performance. This use of machine learning leads to two concerns: (1) The assumption that the training dataset is accurate, where the known error rate for analysts classifying Landsat 8 scenes is about 7% (Foga et al., 2017), produces misclassifications of training data that can skew error assessment and training of the algorithm. (2) The neural networks are difficult to modify for specific use cases that are not covered adequately by the training dataset. For example, the Landsat 8 CCA training dataset has no snow class, nor does it cover mid-latitude mountains, so algorithm performance from these methods over seasonal snow in mountains may be worse.

2.3 Need for an Approach that Combines Spectral and Spatial Information

Multispectral satellites images are different from the RGB images that many deep learning CNN image classification algorithms excel at. The path that radiation takes to hit the multispectral sensor and the spectral properties of the atmosphere and surface of Earth can be physically modeled; algorithms that are based on these processes enable extracting meaningful information from these spectral characteristics. Many classification techniques that work on radiometrically corrected multi-band satellite data do not work on RGB image datasets that CNN semantic segmentation algorithms were designed for.

For many fields, per-pixel classification approaches become challenging when characteristics other than reflectance are the discerning features of the objects, or the objects are far larger than the size of the pixels. Both snowpacks and clouds are larger

than the spatial resolution of the multispectral satellites used to measure them and can have similar spectra. In these cases, approaches to semantic segmentation of images that use techniques other than classical spectral thresholds conceptually resemble the tools humans use when analyzing and masking an image. Multispectral satellite imagery has many additional constraints on image capture that enable the use of methods less obtuse than supervised classification methods for generating cloud masks. In computer vision, often the end goal is the semantic segmentation of the scene; for multispectral satellite imagery, the geolocated, atmospherically corrected imagery allows for the image processing chain to provide physically meaningful interpretation of processes and functions in the Earth system. A transparent, adaptable method that delivers physical properties using the context of that scene can provide additional information that environmental managers, researchers, and scientists can use to learn and understand more about Earth from the imagery. With that goal in mind, the main driving force behind this chapter is to improve spectral thresholding algorithms with an accurate, precise cloud mask that discriminates clouds from snow in ways that are transparent and adjustable by the user.

2.4 Cloud Mask using Illumination, Spectra, and Texture (MIST)

The input data for running *Cloud Mask using Illumination, Spectra, and Texture* (MIST) for Landsat 8 data are the eight 30 m Operational Land Imager (OLI) bands, the 15 m panchromatic band, Shuttle Radar Topography Mission (SRTM) 1 arc-sec void-filled digital elevation model, and the solar position at the time of scene acquisition. The characteristics of the Landsat 8 data required are listed in Table 1. The methods

are readily adaptable to data from Sentinel-2a/b and other multispectral sensors that cover visible through shortwave-infrared wavelengths at spatial resolutions comparable to available digital elevation models.

MIST assumes that snow can vary at a finer spatial scale than clouds and that clouds are larger than the 30 m Landsat pixels. In various mountain ranges around the globe, nearly every MODIS snow covered pixel and 93% of Landsat snow covered pixels show a reflectance signature that indicates the pixel includes reflected solar radiation from different land covers (Selkowitz et al., 2014). In contrast, only around 15% of global cloud cover are from clouds smaller than 10 km^2, where most of these "small" clouds are still much larger than Landsat pixels (Wood & Field, 2011). By assuming all clouds are bigger than Landsat pixels, pixels can be processed with the following two assumptions.

1) Image segmentation techniques that identify objects larger than the size of a pixel are valid approaches for finding clouds in Landsat imagery.
2) Spectral tests and radiative transfer modeling of clouds assume clouds have homogeneous properties at the pixel scale.

MIST has three processing stages. Stage one generates superpixels along with a feature set of characteristics for each superpixel. Stage two uses the feature set for each superpixel to classify it as cloud, cloud shadow, or a clear sky. Stage three merges all the cloud superpixels into a final cloud mask and assigns optical properties to each pixel in the cloud mask. Figure 7 shows a flowchart that gives a detailed overview of the processing steps in MIST.

Superpixels, instead of pixels, are the spatial unit for classification in MIST. This approach increases the likelihood of correctly classifying pixels with ambiguous spectral signatures that are co-located with pixels from the same object that have distinct spectral signatures. Superpixels also provide a useful spatial footprint to apply results from textural and spatial tests for objects where only some individual pixel responses to the textural and spatial tests are indicative and representative of their local region.

2.4.1 Stage One: Segmentation into superpixels with feature sets

Superpixels are an efficient way to create regions within an image that adhere to the boundaries of objects and reduce the number of individual components in the image. Superpixels were developed as a quick and effective method to segment a three band RGB or L*a*b color space image (Achanta et al., 2012). Multispectral sensors have superior spectral information to standard RGB cameras that allow using more than three spectral features from each pixel to create the superpixels. To create superpixels in MIST, the seven 30 m optical bands are used to segment the image into superpixels that are approximately 100 pixels large. In addition, the cirrus band 9 may contain useful information if pixels exceed top-of-atmosphere reflectance values of 0.01 in band 9. At that reflectivity, the pixel may be a cirrus cloud or the land surface at high altitude under a dry atmosphere. When cirrus values exceed 0.01 in an acquisition, the cirrus band is normalized to the 0-1 range to enhance the contrast in the image and scale the values to the range of values in the other bands. This helps with the superpixel segmentation and classification later on in MIST.

Superpixel creation is an iterative process. To initialize superpixel creation, the image is divided evenly, based on image size alone, into X seed regions, consisting of 100-pixels (10x10). The center of each seed region is the initial center location for each superpixel. The initial spectral value for the superpixel is the average reflectance, in each band, of the pixels in the seed region. The superpixels are updated iteratively by comparing each pixel's individual spectra to the mean spectra of every superpixel with a center point less than 100 pixels away. Spectral similarity and spatial proximity are balanced to assign the individual pixel to a superpixel using Equation 2, a spectral angle mapper with the geometric distance as the spectral distance measure and the Euclidian distance between the individual pixel and superpixel center as the spatial distance measure. In a single iteration, each pixel is assigned to a single superpixel based on which superpixel mean spectra their spectra are most similar to and how close the pixel is to the superpixels center. At the end of each iteration, the centers and mean spectra of each superpixel update. After a few iterations, superpixels begin to conform to the shape of objects in the image. Five iterations are used to generate the final superpixels passed to stage two of the algorithm.

$$\boldsymbol{d_{SAM} = cos^{-1}\left(\frac{\overrightarrow{R_1} \cdot \overrightarrow{R_2}}{\left\|\overrightarrow{R_1}\right\|_2 \times \left\|\overrightarrow{R_2}\right\|_2}\right)}$$

$$\boldsymbol{d_{xy} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}} \quad (2)$$

$$\boldsymbol{D = \frac{d_{SAM}}{m} + \frac{d_{xy}}{S}}$$

In Equation 2, $\boldsymbol{R_1}$ and $\boldsymbol{R_2}$ are the vectors of the eight 30 m optical bands for the pixel of interest and the mean of the superpixel. $\boldsymbol{d_{SAM}}$ is the spectral angle between the mean superpixel spectra and the pixel of interest. $\boldsymbol{d_{xy}}$ is the Euclidian distance, in pixel coordinates, between the mean superpixel and the pixel of interest. $\boldsymbol{D}$ is the distance used to assign pixels to a nearby superpixel where $\boldsymbol{m}$ and $\boldsymbol{S}$ are weighting factors to balance the importance of the spatial distance and the spectral distance in assigning pixels to superpixels. The search space for each superpixel is restricted to pixels within $\boldsymbol{2S}$ of the center of the superpixel. Superpixel image segmentation is quick because a pixel is constrained to belonging to a superpixel in its local neighborhood, which greatly reduces the number of pairwise distance calculations needed to assign each pixel to a superpixel.

Once the superpixels are created, a feature set is generated for each superpixel from the spectral, DEM terrain shade, and textural tests in Table 6. These tests are run on the component pixels of each superpixel. The results of the tests are added to the feature set for each superpixel in two ways: either as the fraction of pixels in the superpixel that passed the test, or as a true/false flag, where the result is either true or false for the superpixel. The order in which the tests in Table 6 are used to classify superpixels is shown in the stage two section of the flowchart in Figure 7 and discussed in section 2.4.2.

2.4.1.1 Spectral Tests

The top-of-atmosphere reflectance values in Landsat 8 optical bands are used to calculate spectral indices that help identify pixels and superpixels as cloud or clear sky.

Because the goal of this approach is to excel in snow-covered mountainous regions, thermal bands are excluded because snow and cloud temperatures are not reliably different. Many of the spectral tests are designed to pass pixels that are unlikely to be cloud covered so they may pass some clear sky pixels from land surfaces other than the surfaces they are designed for.

The normalized difference snow index (Dozier, 1989), where $NDSI = (R_{vis} - R_{SWIR})/(R_{vis} + R_{SWIR})$, is used to help identify some snow. Because textural tests are relied upon for identification of lower fractional snow-covered pixels, a conservative snow identification test is used to find pixels with high fractional snow cover that are unlikely to be confused with clouds based on spectra. A thick snowpack with high fractional pixel cover illuminated by direct solar radiation is bright in the visible bands. As a start, to pass the conservative snow test, snow must have a band 2 reflectance above 0.6. This eliminates almost all soils and vegetation and many clouds. Ninety-nine percent of the snow pixels analyzed by Stillinger et al. (2019) had SWIR reflectance below 0.2 in both band 6 and band 7, so for the conservative snow test the pixels must have SWIR band reflectance in both bands below 0.2. Most soils and many clouds have SWIR reflectance higher than 0.2 in at least one of these bands. As a last test to bring in any pixels that do not pass the first two tests, but still have a large difference between their albedo in the visible and SWIR wavelengths, any pixels with an NDSI value above 0.7 are classified as clear sky, this test can also pass some dark land surfaces that also have a high NDSI value. This test is labeled as “snow” in Figure 7 and “snow test” in Table 6.

The whiteness test from Gomez-Chova et al. (2007), used in visible bands from Landsat imagery by FMask (Zhu & Woodcock, 2012), helps identify pixels with flat reflectance in various band combinations. Lower values from this test indicate flatter reflectance across the included bands. In the visible bands this test can separate snow, clouds, whitewater, and sand from most soils and vegetation pixels. A new whiteness test is added to this algorithm which measures the flatness of the reflection for a pixel in the vis-NIR-SWIR. Many cloud pixels are still bright and white in the NIR and SWIR wavelengths, easily seen in a false color image (fRGB) of Landsat 8 Bands 6, 5 and 3. To measure this fRGB whiteness attribute, whiteness is calculated across bands 3, 5 and 6. Pixels that are white and bright in these three bands and have band 2 reflectance above 0.4 are probably clouds. Some soils can pass this test, but most tend to be darkest in the blue wavelengths and significantly brighter in SWIR wavelengths. These two whiteness tests are labeled as W_{432} and W_{653} in Figure 7 and Table 6.

To identify soils, the normalized difference between band 6 and band 1 or band 6 and band 2 must be greater than zero when band 2 has reflectance below 0.25. Band 1 is in a region where soils are quite dark, but atmospheric scattering can increase the measured reflectance in this band. Snow and clouds both tend to be darker in SWIR bands than visible bands so they will have negative index values for these and also tend to have band 2 reflectance above 0.25. This test is labeled as “soil” in Figure 7 and “soil test” in Table 6.

The water test proposed and implemented in Fmask is effective at identifying pixels that contain open water (Zhu & Woodcock, 2012). The water test looks at the relationship between the Normalized Difference Vegetation Index (Rouse et al., 1974),

where $NDVI = (R_{NIR} - R_{red})/(R_{NIR} + R_{red})$, and the band 5 reflectance to identify open water. This test does classify clouds over open water as open water, but the second stage of MIST places these superpixels in the cloud mask with a water flag that uses the median reflectance in band 5 for each superpixel and a cirrus flag to identify clouds over water. These tests are labeled as “water test” and “water flag” in Figure 7 and Table 6.

The cirrus band in Landsat 8 is an excellent tool for identifying high altitude clouds over low elevation terrain. Once elevations get above 3500 m, or the atmosphere is very dry, the land surface can contribute significantly to the signal since snow, vegetation and soils all have surface reflectance between 0.2-0.3 in these wavelengths (Stillinger et al., 2019). A conservative cirrus test accounts for knowledge of the surface elevation of the pixel when determining if the cirrus band can be used confidently to mask cirrus clouds. If the pixel elevation is below 3500 m and the cirrus band reflectance is above 0.05, the pixel is flagged as a cirrus cloud pixel. This test is labeled as “cirrus flag” in Figure 7 and Table 6.

Two vegetation tests are used to extract vegetated pixels from the cloud mask. NDVI values above 0.3 identify heavily vegetated pixels. The dark vegetation test looks for pixels where NDVI is positive, the cirrus flag is false, and band 4 reflectance is below 0.1. The cirrus flag and band 4 threshold reduce the possibility of cloud covered pixels passing this test. This test is labeled as “veg” in Figure 7 and “vegetation test” in Table 6. While called a "vegetation" test, it is likely in many scenes, given the low NDVI threshold, that other land surfaces will pass the test as well. NDVI is being used as an index for identifying cloud free pixels, most of which are vegetated, but in some scenes

will include other land surfaces. Along the same line of reasoning the NDSI thresholds used in the "snow test" pass shaded soils. These are desirable attributes of the spectral tests as these thresholds do not pass cloud covered pixels.

The haze optimized transformation (HOT) was developed to correlate the blue and red band reflectance and detect when additional atmospheric scattering in the blue band, caused by haze and thin clouds, is present (Zhang et al 2002). Bright land cover pixels can pass this test under clear sky conditions. To reduce land cover pixels passing the HOT test and because additional textural tests and the cirrus band tests identify some cloud pixels that the HOT test is designed for, a high threshold of 0.05 is used. This test is labeled as "HOT" in Figure 7 and Table 6.

2.4.1.2 Shade Tests

MIST combines illumination conditions at scene acquisition with the scene topography to include shade and shadow tests for identifying cloudy and clear sky pixels. A set of illumination tests categorizes the shading in the scene to help identify clouds, cloud shadows, and terrain shadows. Superpixel fractional coverage of these categories enables removal of non-cloud superpixels from the cloud mask.

The digital elevation model (DEM) is the version 3 void filled dataset from the interferometric synthetic aperture radar data collected aboard the Shuttle Radar Topography Mission (SRTM: Farr et al., 2007). This elevation data set at 1 arc-sec spatial resolution is globally available between 60N and 56S. Landsat 8 data come with the solar azimuth and the solar zenith angles for the scene center pixel at acquisition time. The 30 m SRTM DEM is used along with the scene center illumination conditions

to calculate per pixel illumination conditions accounting for the surrounding rugged terrain (Dozier & Frew, 1990). Latitude and longitude of each pixel are combined with the elevation grid to calculate local horizons for each pixel. A mask of areas within the image that should be shaded by topography is created by determining whether the sun's position is above or below each pixel's local topographic horizon. Because the spatial scale of the OLI pixels is the same as the DEM, uncertainty in the local horizons makes accurate shadow masks with SRTM data a challenge for a 30 m sensor. Visual inspection of cloud free scenes with terrain shadows and terrain illumination masks created from the SRTM DEM showed false negative bias in the shadows produced by the DEM; not enough shade was in the masks. To adjust for this bias, and still enable the use of 30 m imagery, pixels are classed as sun or shadow pixels given their local horizon in three directions: five degrees to either side of the solar azimuth and at the solar azimuth. For this range of sun azimuths, the shading under both the scene solar zenith and the scene with the sun 10 degrees lower is calculated. Taking the shadow mask generated under these six conditions captures the areas where terrain shadows are expected in the images.

Modeled areas of expected terrain shadows are compared to measured areas of actual shade (OLI shade) to help identify clouds. To measure shade in the optical imagery, the same flood filling technique of the NIR and SWIR bands from CFMask is used to effectively identify local regions of low NIR/SWIR reflectance relative to their surroundings (Zhu et al., 2015). Because atmospheric scattering in the shortwave infrared is less than in the visible part of the spectrum, the illumination of shaded pixels is reduced the most relative to sunny neighboring pixels in the NIR and SWIR

bands. These flood fill steps accurately capture both the shadows in the image, from terrain shadow along with shading from clouds, and also areas that have locally low SWIR/NIR reflectance due to the surface type, like water bodies and snow.

The DEM terrain shadow mask and the OLI shade mask are used in combination to create a possible cloud shadow mask and a possible cloud mask. Pixels that are shadowed in the terrain shadow mask but not the OLI shade mask are flagged as possible clouds, obscuring the terrain shadow that is expected. Pixels that are not shaded in the terrain shadow mask but are shaded in the OLI shade mask are flagged as possible cloud shadows, shade that is not expected given the topography and illumination conditions of the scene. These shade tests are listed in Table 6.

2.4.1.3 Texture Tests

A consistent problem with existing multispectral cloud masks is misclassification of snow as clouds and clouds as snow, due in part to spectral similarity of clouds and snow under some circumstances (Stillinger et al 2019). Fortunately, some of the most spectrally similar snow and clouds are texturally unique. The spatial scale and patterns of spectral variability are different for snow and clouds. Some of these differences are easily seen by humans even in the visible bands of most images, where reflectance differences between snow and cloud pixels are minimal. In mountainous terrain, snow covered pixels with similar reflectance are usually in close spatial proximity to pixels with far higher and lower snow cover fractions and significantly different reflectance values. This range of snow pixel reflectances over a small area creates a characteristic texture for some types of snow cover that is distinct

from the homogeneous texture of clouds. Clouds tend to be much larger that the size of a 30 m Landsat pixel and have a smooth texture, with consistent reflectance values in all bands across large areas. Additionally, the patterns of shade, shapes of the clouds, and extent of their cover are not constrained to match the topography underneath them. Mountainous terrain exhibits a characteristic fractal texture, with patterns of ridges and valleys expanding from peaks and alternating patterns of lighter and darker mountain sides, based on their slope, aspect, cover type, and illumination. The texture filters in this section encode the information for a local region of pixels into values that discriminate superpixels between cloud and clear sky. Two filters were developed, one identifies spectrally ambiguous snow based on its textural pattern that is unique from cloud texture, the second flags regions of OLI images where expected topographic patterns, given the local illumination conditions, are missing in the OLI acquisition. These textural differences are used to filter out clear sky pixels from the possible cloud pixels.

Gabor filters are used to identify snow that cannot be cloud based on texture. First used to process 1D information signals (Gabor, 1946), later they were expanded to the spatial domain and shown to be an accurate model for the low level vision responses for cortical cells in mammals (Hubel & Wiesel, 1965; Pollen & Ronner, 1983). Soon afterward, Gabor filters were proven to function well on natural images to extract texture (Daugman, 1985; Field, 1987). They have since been heavily investigated for applications in texture identification and applied to a range of problems from defect detection in automatic textile production (Kumar & Pang, 2002) to fingerprint recognition, generalized object detection, and semantic segmentation of

images (Jain et al., 1997). For snow-cloud discrimination, a Gabor filter bank is tuned to respond to large amplitude, short wavelength, repeating texture in the panchromatic band. The 15-m panchromatic band (band 8) is used because the spatial variability of snow extends to the sub-15 m scale, whereas clouds tend to be homogeneous over large, multi-pixel regions at this scale. Using the panchromatic band enables a high confidence threshold to be set that ensures no cloud pixels pass though the filter. The Gabor filter bank used in MIST is described in the last row of Table 6 and is labeled as "Gabor" in Figure 7.

A Gabor filter selects for frequency information in the image based on four tunable parameters: the wavelength in pixels per cycle of the repeating texture, filter orientation relative to the image coordinates, spatial-frequency bandwidth, and spatial aspect ratio. To tune the filter to respond to the rapid spatial changes in spectral variability of mountain snow cover, the Gabor filter bank is set with a fixed wavelength of 3 pixels per cycle. The orientation of the filter is the orientation in the image plane that the filter is sensitive to the frequency information. The spatial-frequency bandwidth controls how much deviation from the preferred frequency wavelength of 3 pixels/cycle is allowed to pass through the filter, it is set to 1 octave which prevents the filter from allowing any textural patterns that repeat across scales larger than a few 15m pixel to pass through the filter. The spatial aspect ratio controls the shape of the filter in the spatial domain, setting this parameter to 0.5, the filters response decay's the same in all directions.

Gabor filter banks traditionally include many filters set to various tunings, generally done to enable the filter bank to identify a range of unknown objects with

various textures. For this algorithm, the sole function of the Gabor filter bank is to respond to mountain snow and a single tuning is sufficient.

The panchromatic band is convolved with Gabor filters oriented at 0°, 45°, 90° and 135° relative to the image coordinates, to make the filter response rotationally invariant to the image so that at least one of the filters will capture the texture, no matter its orientation in the image. The response when filter banks exceeded 4 orientations has minimal impact on classification results (Bianconi & Fernandez, 2007), so only 4 orientations are used to reduce processing requirements of the filter. The sum of squares is calculated on the 3-D filter bank output for each 15m pixel to reduce the data from the four filter orientations into a 2D image and increase the differences in value between high and low filter responses. Then, the maximum value from this 2D image within the 2x2 pixel footprint of the 15m panchromatic pixel overlain on a single 30 m OLI pixel, is assigned to the 30m OLI pixel and used to determine if the 30m OLI pixel passes through the Gabor filter bank. Choosing a maximum value from the 16 Gabor filter values within a 30 m pixel (four filter orientations per pixel and four 15m pixels within one 30m pixel) with this method increases the filter response differences between clouds and snow. Most cloud pixels will have low responses in all 16 values and a snow-covered area may only have a few values that exhibit a strong response to the filter.

To set the threshold above which pixels are classified cloud free, the Landsat 8 CCA validation cloud masks are used to test the cloud pixels response to various tunings of the Gabor filter bank (Foga et al., 2017). These validation masks are 25 full Landsat scenes that contain snow, but have no snow mask, so they are not used in

algorithm error assessment and offer an independent way to set the Gabor filter bank cutoff value. Figure 8 shows the various thresholds applied to the Gabor filter bank output plotted against the fraction of cloud pixels in each of the 25 scenes that pass through the filter at that cutoff value. The goal is the lowest filter bank response threshold that minimizes cloud pixel inclusion. The Gabor filter threshold value was set to 0.25 based on the results in Figure 8. To verify snow pixels are not excluded by the filter cutoff, the SPARCS USGS scenes were passed through the filter and the median fraction that pass for both snow and cloud was calculated. Figure 9 shows that 0.25 is an excellent cutoff point as minimal cloud pixels pass the filter, but most of the snow pixels still generate a filter response.

The second texture test is a terrain based approach that flags areas where expected topographic patterns, given the local illumination conditions, are missing in the OLI acquisition. Histograms of Oriented Gradients (HOGs) are compared between modeled hillslope illumination at scene acquisition and the OLI acquisition, to flag when lighting gradients and orientations in the hillslope image are present in the OLI image. When clouds pass over mountainous terrain, they obscure the terrain and areas where HOGs do not match between the hillslope image and OLI image may be clouds.

HOG was originally developed for quick, automatic pedestrian detection in images and video (Dalal & Triggs, 2005). The original goal was to create a detector that was sensitive to the characteristic outline of upright humans, and could be used alongside a support vector machine classifier to determine if full images or image regions had a human present. The use of HOG in MIST is different, the same methods for calculating HOG are used, but instead of comparing the OLI acquisition to a library

of labeled training data, each OLI region is compared to the same location in the illuminated hillslope with topographic horizons created from the DEM. HOG creates a vector that is a representation of the gradient magnitudes and directions in an image. Because the hillslope and clear sky OLI have similar light to dark and dark to light transitions in mountainous terrain, similar HOG feature sets in each image only occur when the terrain is not obscured by clouds in the OLI image. The OLI acquisition is only compared, region by region to the corresponding location in the hillslope image to find matching areas, defined as areas where the difference between HOGs are less than 0.9. Because clouds obscure the terrain, and do not follow the same illumination patters, missed matches can be flagged as possibly cloudy areas. The HOG test used in MIST is labeled as "HOG flag" in Table 6 and Figure 7.

Some snowy areas and other regions where the hillslope illumination does not closely resemble the gradients in the actual spectral reflectance of the hillslope are also missed, but do contain information from other tests in MIST to help class the superpixels they are within. The lack of distinguishing topographic or spectral features known with high confidence to be land features is a main mechanism MIST uses to find cloud covered areas in OLI imagery.

2.4.2 Stage Two: Classify superpixels

The second stage of MIST takes the feature set of spectral, shade, and texture features for each superpixel and uses the feature set in a decision tree to classify each superpixel as cloud, cloud shadow, or clear sky (Figure 7). At the start of the decision tree, all the superpixels are assumed cloud covered. MIST then searches for

characteristic features of clear sky pixels to remove superpixels from the cloud mask. Two types of features in the feature sets are used to classify the superpixels: fractional cover features and flag features, Table 6 lists the category for each feature. Fractional cover features count the fraction of pixels in the superpixel that pass that test in stage one. Flag features are true or false for the superpixel as a whole, and can be set by just a single pixel within the superpixel passing that test. This method of classifying pixels via superpixel characteristics enables pixels that are adjacent to known cloud pixels, and likely part of the cloud due to their inclusion in the superpixel, to remain in the cloud mask even if they do not individually pass tests to be considered a cloud pixel. Figure 7 shows the flowchart of stage two and the way the features in Table 6 are used to classify the superpixels.

If the superpixel has some characteristics of the land surface, it is passed through a set of tests to find thin clouds and separate thin clouds from clear sky. If no evidence of land cover is found, but the superpixel is possibly a cloud shadow, it is tested to separated cloud on cloud shadow from a shaded land surface. If no tests are passed, the superpixel is classed as a cloud covered pixel.

2.4.3 Stage Three: Assign optical properties to cloud mask pixels

This mask, unlike other Landsat 8 cloud masks, also estimates the clouds' optical properties. Knowing the phase, particle size, and water equivalent of the cloud helps with estimation of the transmission of shortwave and emission of longwave radiation from the cloud to the surface below.

The radiative transfer model described by Stillinger et al. (2019) is used to create a lookup table of modeled cloud reflectance across the full range of expected cloud phases, thicknesses, droplet sizes, and underlying cover. The analysis includes clouds of varying optical thicknesses over different land surfaces. Reflectance of cloud pixels in sensor band passes depend on the size of the ice crystal or water droplet, the phase of water (liquid or solid), the optical depth of the cloud, and the reflectance of the underlying surface. Cloud spectra are modeled as homologous mixtures of liquid droplets from 5 μm to 25 μm with water equivalents from 0.01 to 10 mm, or ice spheres from 5 μm to 50 μm with water equivalents from 0.01 to 3 mm. All cloud types are modeled with a surface reflectance of conifer vegetation and again with bare soil.

Optical properties of the pixels in the cloud mask are estimated by matching the pixels measured reflectance to the closest matching cloud spectra from the reflectance library. Spectral angle mapper is used to match pixels to reference spectra from the library and the parameters of the library spectra are assigned to the cloud pixel. The phase of the cloud, its radius, water equivalent, and wetness are estimated. Inverting each pixel's optical properties could provide more precise estimates, but attempts at inverting many pixels per sub scene was untenably slow. Using the lookup table provides coarser estimates of the clouds' optical properties but far quicker.

2.5 Accuracy Assessment

The performance of MIST was compared against manual masks from OLI. The MIST cloud mask classification was compared to each pixel's manual classification. Each pixel was categorized as:

True Positive – TP: Both MIST and analysts classify the pixel as the class of interest.

True Negative – TN: Both MIST and analysts classify the pixel as not the class of interest. The classification of the analyst and the automated mask do not have to match.

False Negative – FN: Analysts classify the pixel as the class of interest, and MIST classifies it differently.

False Positive – FP: MIST classifies a pixel as the class of interest, and the analysts classify it differently.

For each scene and for all scenes aggregated together, the total counts of each category (TP, FP, TN, FP) for the cloud mask are totaled. The following metrics are used in the error assessment of MIST (Olson & Delen, 2008):

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F = 2\frac{Precision \times Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Precision represents the probability that a pixel identified as a class is indeed that class. *Recall* represents the probability of detection of a pixel of the class of interest. The *F* score balances the two metrics.

The validation dataset comprises 10 scenes from the 2014 USGS SPARCS manual classification dataset (U.S. Geological Survey, 2016) of Landsat 8 OLI sub-scenes of 1000×1000 pixels, with each pixel identified as snow, cloud, cloud shadow,

water, land, or bad data. Of 80 scenes in this dataset, 13 included significant snow within the manual masks and 10 of these scenes overlapped with the SRTM DEM dataset required for MIST to operate.

2.6 Results

MIST is able to mask cloud covered areas in Landsat 8 images with high precision and moderate recall. Table 7 shows MIST results alongside the CFMask cloud mask in the BQA file of the same image. The overall and scene-by-scene precision of MIST is greater than CFMask. Overall recall is about the same, whereas scene-by-scene Recall of MIST is slightly better than it is for CFMask. Overall, when looking at the F statistic, MIST offers a 10% improvement over CFMask in mountainous terrain. There are specific circumstances where MIST has difficulty. The first is an issue misclassifying some land surface features that cover many superpixels and have homogeneous texture. In Figure 13 some large soil covered regions in the mountain valleys are misclassified as clouds. In Figures 13-17 clouds smaller than the size of a superpixel are not classified as clouds by the algorithm. Some superpixels along both the boundary and in the central interior of clouds are sometimes not classified as clouds by MIST. While the recall rate for MIST is similar to past algorithms, it appears that the False Negative cloud pixels in MIST are different pixels than the False Negative cloud pixels from prior operational masks.

Table 8 provides error statistics for all 10 scenes, along with each acquisition's terrain characterized by the percent cover of mountain landforms from the global Hammond landform regions (Karagulle et al., 2017), which quantifies Earth's

mountainous locations at 250 m horizontal scale. In 8 of the 10 scenes, MIST has superior precision compared to CFMask, in 4 of 10 scenes MIST has superior recall. Performance for all scenes are similar for MIST with the exception of WRS P170 R25 from March 21, 2014 where MIST does not perform adequately. CFMask also has the worst performance out of all scenes on this scene.

Figure 11 and Figure 12 show the ability of the superpixel algorithm to adhere to the boundaries of clouds, shaded terrain, and snow. Depending on the parametrization of the weighting between spatial and spectral distances, the superpixels are either compact or have irregular boundaries. From these figures it can be seen that areas inside the middle of snowpacks and large clouds have superpixels that are hexagonal shape showing the maximal extent they reach before the distance weighting classes the pixel in the adjacent superpixel. Figure 14 and Figure 15 show the output of the optical properties from the cloud mask. Distinct connected regions within the clouds are estimated to have similar optical properties with banding patterns indicating as cloud top heights vary through the clouds, the portions of the cloud top at the same altitude in the atmosphere may have similar optical properties. A combination of mixed phase, water, and ice clouds are found to best match the optical properties measured in the test scenes. Large portions of the cloud masks are modeled with cloud particle ice radius values from 10-40μm, appropriate for the particle range expected in clouds and unrealistic for a mountain snowpack.

2.7 Discussion

Cloud masking using spectral, shade, and textural tests to classify superpixels generates precise cloud masks over snow covered mountain ranges.

Compared to prior cloud masks assessed for Landsat 8, MIST exhibits an increase in performance in terms of precision with no reduction in recall. Table 9 shows MIST compared to the results of prior error assessments for operational Landsat cloud masks globally and over snow covered mountain ranges. MIST precision over mountainous terrain is similar to the global performance of CFMask, whereas CFMask performance is worse over mountainous terrain compared to MIST. MIST incorporates techniques that improve precision over snowy mountains that the operational product currently lacks.

Superpixels offer a simple way to efficiently use local contextual information for making classification decisions. Superpixels ability to follow the boundaries of objects, including clouds over snow, and shaded terrain, lead to their usefulness as a replacement for the pixel as the unit of analysis for making classification decisions. By assigning each superpixel a set of characteristics based on the features of their component pixels, richer information can be used to help class the superpixel. This approach helps in areas where the objects are much larger than the pixel size, and significant sub-object spectral variability exists, and it also helps in cases where objects contain pixels that cannot be classified with high confidence based on spectral signature. Superpixel generation is easy to modify by changing the initial seed locations of the pixels, the features used for each pixel, the size they are allowed to grow to, constraints on their shape, and how they can be merged into larger objects.

Combining the flexibility of superpixel creation with the ability to adjust the composition of the feature sets and how they are used to classify the superpixels increases the flexibility in how masks are created and the adaptability to other remote sensing applications apart from cloud masking.

Prior studies have used superpixels to create Landsat cloud masks. They used three bands and a Euclidian distance measure for spectral distance to create the superpixels, adapting the algorithm developed for RGB images without altering the foundational mechanism of superpixel generation to take advantage of the 8 calibrated spectral bands in Landsat 8. Additionally, after superpixel creation, these prior masks mainly relied on machine learning approaches to class superpixels as cloud covered or clear sky. Existing manual masks were used as training data to class clouds. MIST retains the benefits of transparent and adjustable spectral tests, and includes new spatial information tests as well, which allow the user to mask clouds without the need to train a machine learning algorithm. By using spectral angle mapper with 8 bands to generate the spectral distance for superpixel creation, the ability to group pixels into superpixels effectively is improved.

Prior uses of terrain data for creating Landsat cloud masks have been limited and not used the full suite of data available from a DEM. FMask can use the terrain to adjust its estimate of where a cloud shadow will be projected, but does not include any additional uses of the terrain data for improving cloud masking. Illumination conditions over mountainous terrain can inform more than just the location of the shadows. By accounting for the patterns of illumination that a hill shade image can provide, clouds can also be identified as areas where this expected pattern is missing.

The Histograms of Oriented Gradients used in MIST is an example of how the hill shade information can be used and an application of computer vision algorithms to multispectral satellite images. The HOG method used in this paper was developed to output a single feature set vector for a whole image and for that vector to be passed through a SVM (support vector machine) classifier to determine if there was a human anywhere in the image. The application of HOGs in this paper uses the same calculation of Histograms of Oriented Gradients, but instead of appending all HOGs into a single feature vector, MIST constructs multiple feature sets by considering the individual HOG calculation for each superpixel as an individual "image." Then, instead of comparing to a library of training data, the OLI HOG feature set for each individual superpixel is only compared to the known hillslope illumination of the superpixel at the time the picture was taken. This method is designed to replicate a visual check an analyst would conduct between what mountains are known to look like and what is scene in the image. Instead of or in addition to a hillslope shade image, prior clear sky acquisitions of the same area could be used in a similar fashion with the same test to see if the terrain illumination and spectral composition of the region is as we would expect in the OLI image.

A main contribution of this chapter is the discovery of the ability to tune Gabor filters to discriminate spectrally similar snow and clouds. These tuned Gabor filters work at identifying snow but not clouds because of the difference between the textures of snow and cloud. In the past, Gabor filters were used on AVHRR remote sensing data for cloud masking and cloud type identification, but the filters were tuned to respond to cloud pixels, not snow (Lamei et al., 1994). This new approach works because the

Gabor filters are selected to respond to non-cloud pixels that have similar spectral characteristics to cloud. This approach takes advantage of the fact that clouds are homogeneous, large, "smooth" areas of texture and will show up as large areas of minimal filter response.

Computer vision techniques, like HOG and the Gabor filters used in MIST, still have unexplored potential for improving classification of multispectral satellite data. Commercial development and applications of computer vision has exploded in recent years, ranging from face recognition software for unlocking cell phones, to real time semantic image segmentation of images for self-driving cars. Many of the use cases that drive development of new methods are for images taken in difficult conditions. Camera position and illumination conditions can freely vary from image to image, the actual cameras vary and only have RGB sensors, and results might be needed in real time. These difficult conditions have resulted in the development of robust methods to class objects that have not yet been explored for their full applicability to multi spectral satellite images. From the perspective of implementing these computer vision tools, multispectral satellites are a simpler application of these tools. Landsat and other satellite sensors repeatedly image the same location with radiometrically corrected multi band images, from a fixed viewpoint, with known camera position, and known illumination and topography. MIST shows that computer vision tools like Gabor filters and HOGs combined with superpixels can fine tune older thresholding algorithms and improve our ability to extract information from earth observing satellites without resulting to black box machine learning methods.

While MIST offers a significant improvement in precision, the current version does not improve the recall of cloud masks in OLI images. MIST struggles with correctly identifying clouds that are significantly smaller than 100 pixels in size, as the superpixel segmentation does not have the ability to find these small objects and separate them from their surrounding pixels as the superpixel generation algorithm tends to generate superpixels that are 100 pixels large due to the spacing between superpixels and distance weighting factor used to control superpixel growth. A fallback per pixel method to try and find these very small clouds would improve MIST if it could be constructed to not include clear sky pixels. Additionally, MIST has a hard time with the edges of cloud objects in some situations and does not correctly add cloud edges to the cloud mask. Some of these edges are thinner portions of the clouds, which are known difficult regions to classify.

A future developmental step of MIST is to include a snow masking step whereby existing spectral unmixing of fractional snow cover assessment methods are used in a second pass after MIST has masked all the clouds. This second pass could identify lingering cloud pixels that are unmixed as snow with too fine a grain size, and also add a useful snow mask to the output. The spatial patterns in RMS error in the mixing model in the second pass could find small clouds and thin cloud boarders. By intentionally not including a cloud end member, sub-pixel or thin clouds may be unmixed with a high RMS. The spatial pattern of the RMS could represent the border of the cloud. If a scene is cloud free, or cloud pixel locations are known with high recall and precision, then spectral unmixing of fractional snow cover can produce an

excellent snow mask that can detect smaller fractions of snow cover per pixel that the existing snow mask for CFMASK.

Future applications of MIST include trying to expand its use to sensors at varying spatial and spectral scales. Cloud masks for MODIS, with their high spectral resolution but lower spatial resolution, may be improvable by including superpixels and textural tests in the processing workflow. Another dimension in which superpixels may work well, that has not been explored with satellite data, is in the time dimension. Snow cover would maintain a monotonic boundary in the time dimension during the melt season whereas clouds may be identifiable by rapid changes in the shape of the superpixels for a day or two. It is not apparent without investigation whether snow will exhibit the necessary textural response at the coarser spatial scale for Gabor filters and HOG to have the same effect. Luckily only a single band at high spatial resolution is needed to create the filter bank, so sensors like GOES, MODIS, and VIIRS, which have a few higher spatial resolution bands at 500 m, 375m, and 250m respectively, may benefit from the Gabor filter bank and HOG tests developed here. If they do, that would bode well for improved snow cloud discrimination at that scale.

2.8 Conclusion

This algorithm developed a texture identification method for mountain snow and extends the use of superpixels to multidimensional space with Spectral Angle Mapper. Gabor filters were shown to be an effective method of discriminating spectrally similar snow and cloud that is texturally distinct and HOG was shown to work at identifying cloud free areas by matching expected and observed terrain

features. This chapter touches on some of the ways cloud masks can be improved without resorting to machine learning techniques that fail to provide traceable reasons behind classification decisions. Improved cloud masks enable accurate use of snow measurements to forecast water supplies and reconstruct snow water equivalent. Considering performance of operational masks and developing new cloud masks that function well in mountainous snow-covered terrain is an important application of multispectral satellites. By improving the usability of multispectral satellite data, our ability to measure and manage our global water resources improves. Mountains only cover a quarter of Earth's surface and are not a standard environment for validating operational cloud masks. While they only cover a fraction of the globe, their snow and ice comprise the main water supply for 2 billion people, and remote sensing tools are a main resource for water supply management in these regions.

2.9 Tables

Table 6 List of spectral, shade, and texture tests used in MIST. The category of the test, name of the test, equation, type of test and reference are included. Tests are either fractional tests - the fraction of pixels in a superpixel that pass the test is the number added to the feature set, or the test is a flag test –if at least a single pixel within the superpixel passes the test that flag is added to the superpixel feature set

	Test Name	Equation	Superpixel Test Type	Reference
Spectral Tests	Normalized Difference Snow Index (NDSI)	(B3-B6)/(B3+B6)	intermediate step	Dozier 1989
	Normalized Difference Vegetation Index (NDVI)	(B5-B4)/(B5+B4)	intermediate step	Rouse et al. 1974
	Haze Optimized Tranformation (HOT)	(B2-0.5B4-0.08) >0.05	fractional	Zhang et al. 2002
	visable band whitness (W_{432})	μ_{432} = (B2+B3+B4)/3 W_{432} = (|B2-μ_{432}|+|B3-μ_{432}|+|B4-μ_{432}|)/μ_{432}	sort and filter	Gomez-Chova et al. 2007
	VIS-NIR-SWIR whiteness (W_{653})	μ_{653} = (B3+B5+B6)/3 W_{653} = (|B3-μ_{653}|+|B5-μ_{653}|+|B6-μ_{653}|)/μ_{653} W_{653} = W_{653} <0.7 & B2>0.4	fractional	this paper
	cirrus flag	(DEM<3500m) & (B9>0.05)	flag	this paper
	snow test	(B2>0.6&B6<0.2&B7<0.2) | NDSI>0.7	fractional	this paper
	soil test	[(B6-B1)/(B6+B1) >0 | (B6-B2)/(B6+B2) >0] & B2<0.3	fractional	this paper
	water test	(NDVI<0.01 & B5<0.11)| (0<NDVI<0.1 & B5<0.05)	fractional	Zhu & Woodcock 2012
	water flag	median($B5_{superpixel}$) <0.1	flag	this paper
	vegetation test	NDVI>0.3	frational	
	dark vegetation test	NDVI>0 & B4<0.1 & ~cirrus test	fractional	this paper
Shade Tests	OLI shade	floodFill = minimum([floodFill(B5)-B5] | [floodFill(B6) - B6]) OLI shade = floodFill >0.02 & ~waterTest	intermediate step	Zhu & Woodcock 2012
	terrain shade	local topographic horizon > solar zeinith angle	intermediate step	Dozier and Frew 1990
	possible cloud shadow	~terrain shade & OLI shade	fractional	this paper
	possible cloud	terrain shade & ~ OLI shade	fractional	this paper
Texture Tests	HOG flag - Histograms of Oriented Gradients (HOG) comparisons with Local Zenith Angle (LZA) and OLI image	euclidainDistance[$LZA_{superpixel}$(HOG),$OLI_{superpixel}$(HOG)]<0.9	flag	this paper
	clear sky panchromatic (B8) gabor filter	orientations: 0° 45° 90° 135° (4 band filter) wavelength: 3 pixels spatial frequency bandwith: 1 spatial aspect ratio: 0.5 stretchlim(max(L2norm(4 band gabor filter), 2x2 window)) >0.25	fractional	this paper

Table 7 Error statistics for MIST compared to CFMask

	Landsat 8 cloudMIST			Landsat 8 CFMask			
Aggregated scenes				Aggregated scenes			
	Precision	*Recall*	*F statistic*		*Precision*	*Recall*	*F statistic*
Cloud	0.916	0.834	0.873	Cloud	0.695	0.857	0.768
Scene-by-scene	median (min, max)			Scene-by-scene	median (min, max)		
	Precision	*Recall*	*F statistic*		*Precision*	*Recall*	*F statistic*
Cloud	0.909 (0.35,0.95)	0.833 (0.50,0.92)	0.874 (0.41,0.92)	Cloud	0.803 (0.03, 0.99)	0.793 (0.03, 0.98)	0.771 (0.03, 0.96)

Table 8 Scene-by-scene comparison of MIST results and CFMask results for cloud masking in the 10 validation scenes.

Date	Sensor/ Product	Scene Coordinates	Mountain Cover	Geographic Coordinates		*Precision*	*Recall*	*F statistic*
13-Jan-2014	CFMask	WRS P 76 R 91	67%	-45	168.3	0.765	0.892	0.824
	cloudMIST					0.904	0.848	0.875
15-Apr-2014	CFMask	WRS P 137 R 41	90%	27.8	90.8	0.856	0.965	0.907
	cloudMIST					0.951	0.897	0.924
5-Jun-2013	CFMask	WRS P 195 R 28	89%	46.4	7.8	0.766	0.939	0.844
	cloudMIST					0.930	0.659	0.771
1-Apr-2014	CFMask	WRS P 167 R 51	60%	47.8	-113.9	0.992	0.801	0.887
	cloudMIST					0.914	0.880	0.897
10-Dec-2013	CFMask	WRS P 231 R 94	80%	30.4	78.2	0.778	0.894	0.832
	cloudMIST					0.945	0.901	0.923
14-Aug-2014	CFMask	WRS P 47 R 25	98%	33.8	74.3	0.792	0.953	0.865
	cloudMIST					0.934	0.817	0.872
14-Jul-2013	CFMask	WRS P 148 R 35	100%	64.8	-17.0	0.865	0.780	0.820
	cloudMIST					0.900	0.817	0.857
21-Mar-2014	CFMask	WRS P 170 R 25	0%	38.4	-119.4	0.685	0.670	0.677
	cloudMIST					0.346	0.504	0.410
26-May-2014	CFMask	WRS P 15 R 24	8%	63.7	-49.3	0.706	0.906	0.794
	cloudMIST					0.865	0.916	0.890
28-Dec-2013	CFMask	WRS P 149 R 37	67%	-49.1	-73.0	0.807	0.852	0.829
	cloudMIST					0.903	0.737	0.812

Table 9. Comparison of MIST results to previously published cloud masking algorithms

	Precision	Recall
cloudMIST (this paper)	0.916	0.834
CFMask for clouds in snow-covered mountains (Stillinger et al 2019)	0.695	0.857
CFMask for snow and ice (Foga et al., 2017, Table 7)	0.757	0.866
CFMask global assessment (Foga et al., 2017, Table 5)	0.880	0.973
Fmask global sample (Hughes & Hayes, 2014)	0.740	0.986
ACCA "best cloud mask over snow/ice" (Foga et al., 2017, Table 7)	0.706	0.993

2.10 Figures

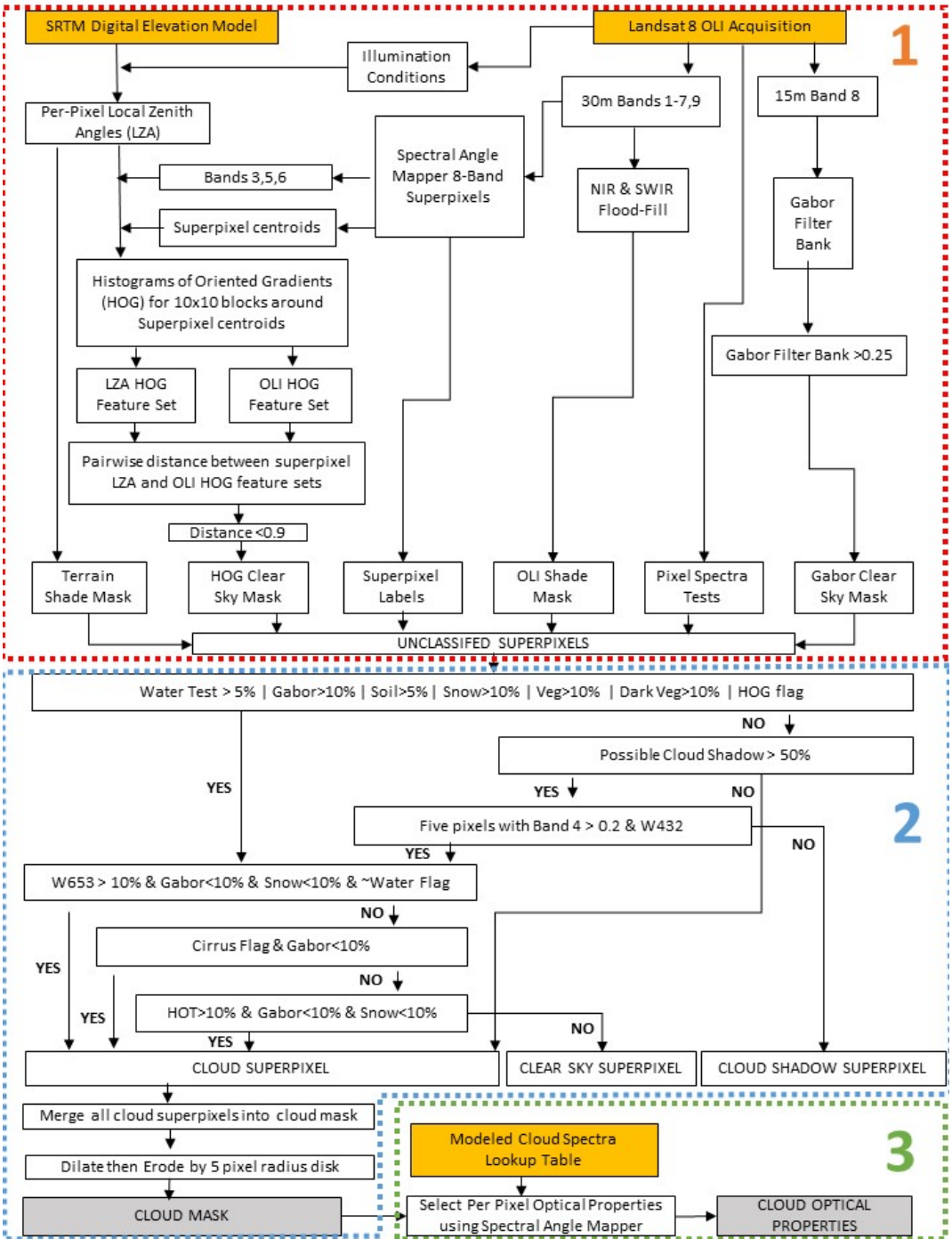

Figure 7. MIST flowchart. The algorithm is divided into three stages, identified by matching number and color in the figure. Stage one: Creating superpixels with feature sets. Stage two: Classifying superpixels. Stage Three: Estimating optical properties from the cloud mask. In stage two, fractional

cover tests count the fraction of pixels in the superpixel that pass the test. For example, Soil>5% means at least 6 pixels in a 100 pixel superpixel would need to pass the soil test in Table 6 to send that superpixel down the "yes" side of the first decision tree branch in stage two. Flag features are true or false for the superpixel as a whole, and can be set by just a single pixel within the superpixel passing that test. The names of the individual spectral, shade, and textural tests and the calculations they represent are listen in Table 6.

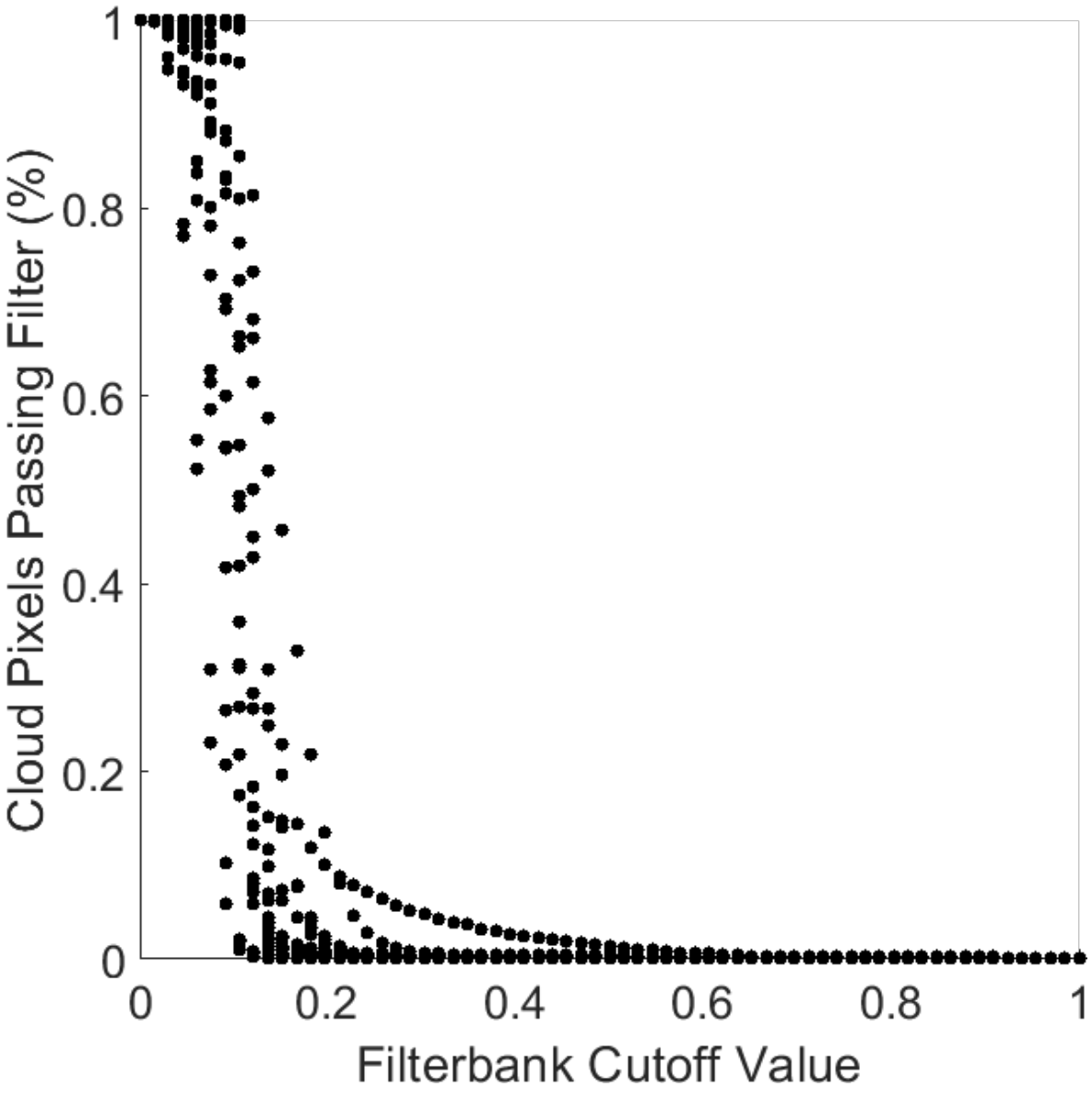

Figure 8: Panchromatic filter bank response to cloud pixels from 25 Landsat 8 scenes. Minimal numbers of cloud pixels pass the filter bank cutoff value of 0.25. This graph was used to set the filter bank response threshold value.

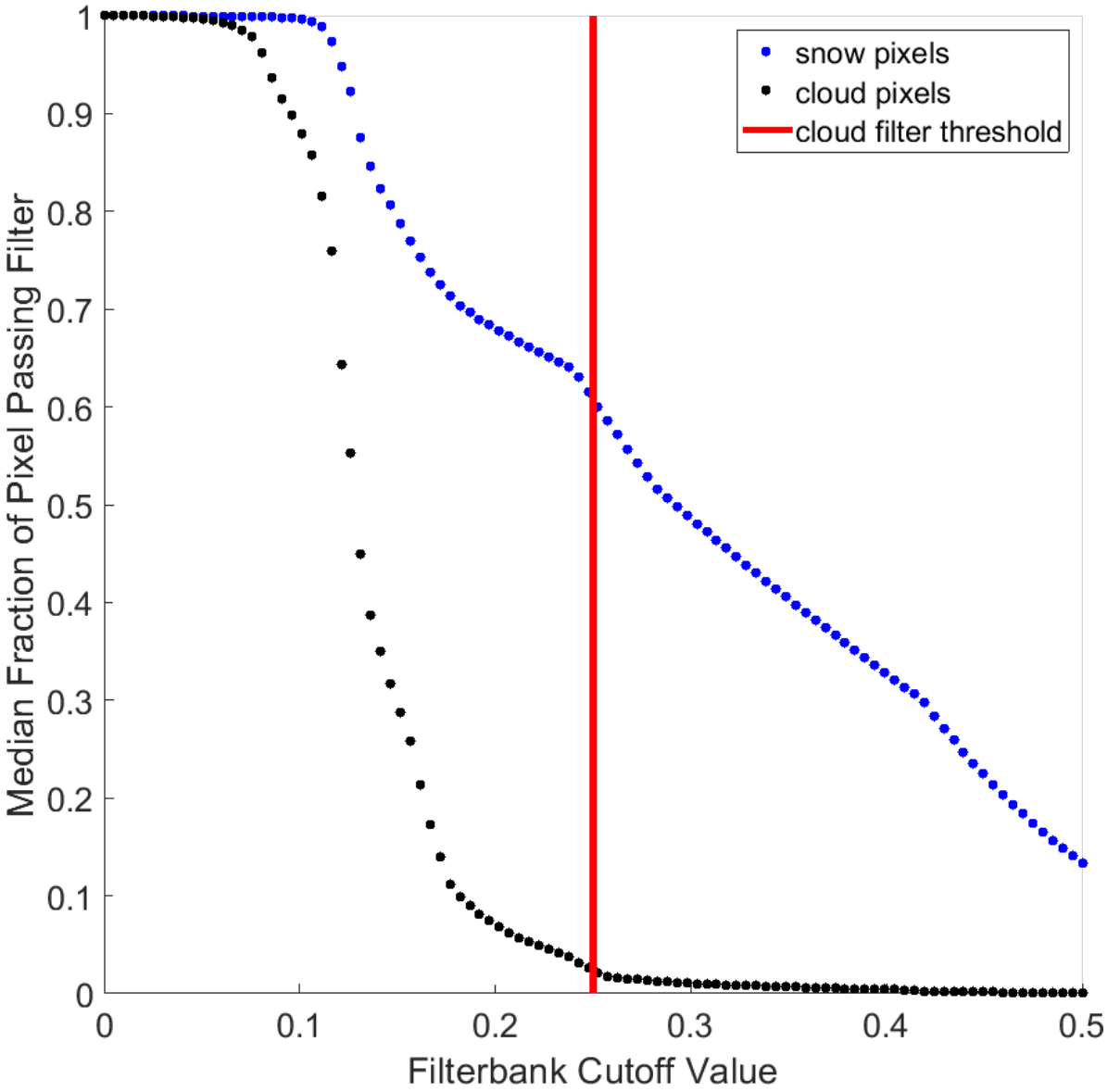

Figure 9: Gabor filter bank cutoff value with per scene median fraction of snow and cloud pixels passing the filter for 13 Landsat scenes. Many snow pixels pass the filter bank cutoff while cloud pixels do not. The filter can effectively only let non-cloud pixels through.

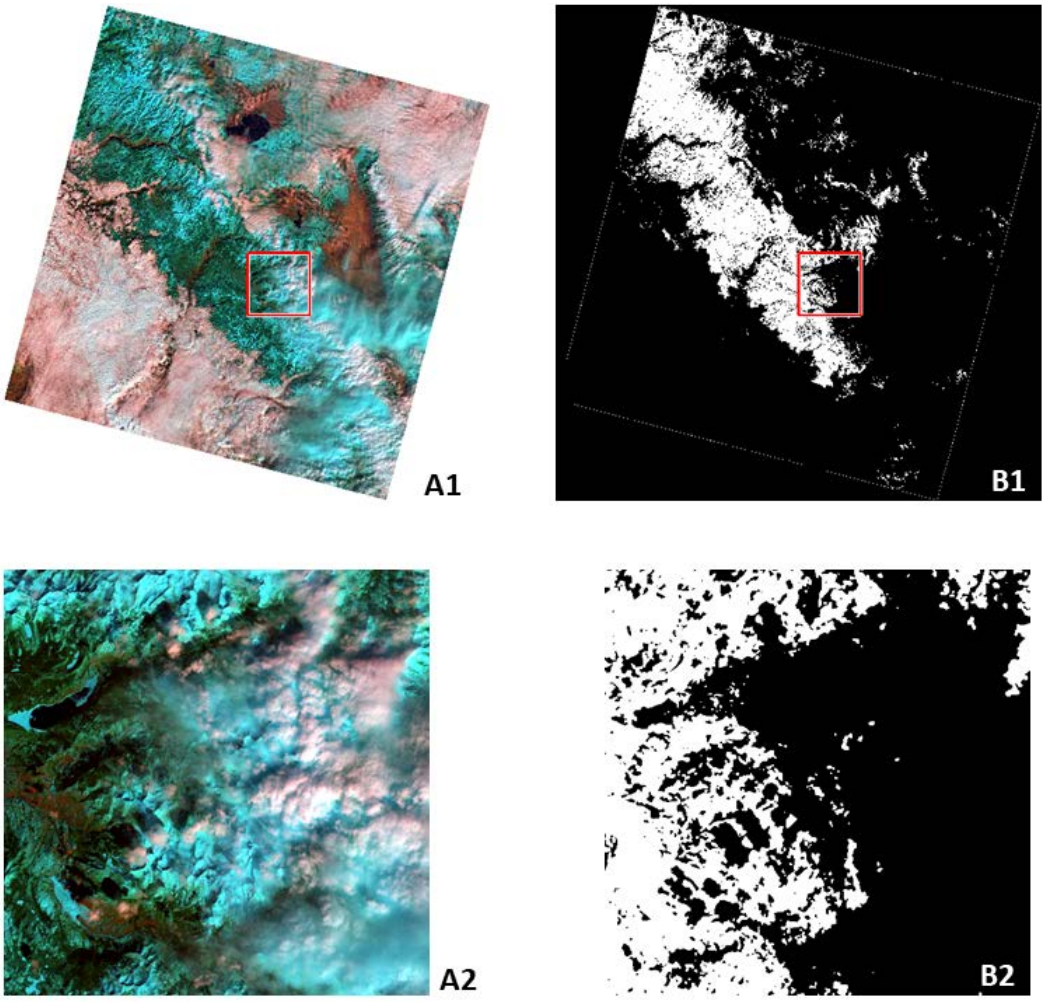

Figure 10: Example of the performance of the panchromatic Gabor filter bank in an image of snow and clouds. The top row shows images of WRS path 42 row 34 over the Sierra Nevada with a red region-of-interest for the zoomed in panel on row 2. A1 and A2 images are false color RGB images of bands 6, 5, and 4. B1 and B2 images are binary masks of the panchromatic Gabor filter bank above a threshold value of 0.25. Clouds, water bodies, and some shadows are vast areas not in the mask. Most snow-covered areas are a mix of many masked pixels with some unmasked pixels. Using a superpixels fractional cover of masked pixels can effectively discriminate between clouds and snow in mountainous terrain.

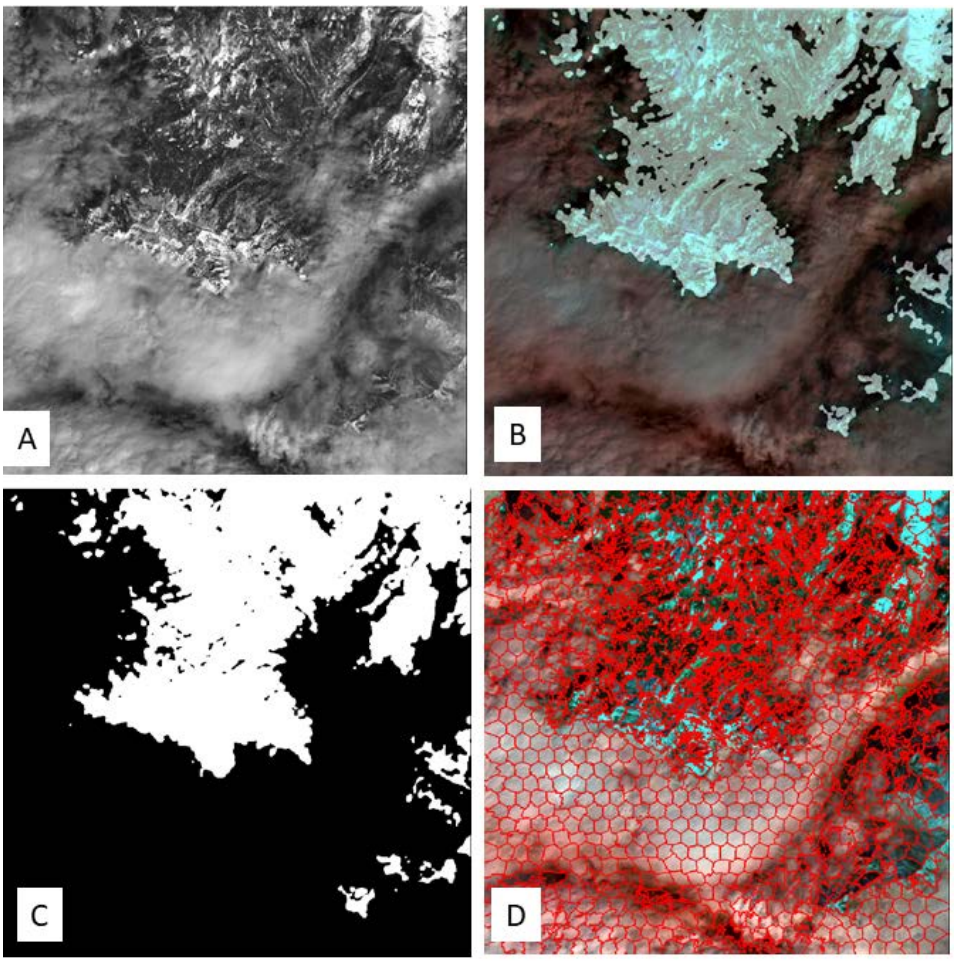

Figure 11: Example of Gabor filter bank output next to superpixels. (A)15m panchromatic band acquisition of an area of snow and cloud color with similar reflectance values in the panchromatic band. (B) Overlay of the cloud free mask generated by only a Gabor filter bank. (C) Binary mask that show cloud free pixels that pass the Gabor filer bank. (D) Superpixel boundaries that are used with the Gabor filter bank output to identify cloud free regions. Cloud covered areas are vast expanses of pixels that do not pass through the filter. Superpixels of cloud cover will have less than 10% coverage (usually 0%) of pixels that pass the filter. Snow that does not pass the filter cutoff, is usually surrounded by snow pixels that do, so the superpixels that contain these false negative snow pixels will still be classed as snow.

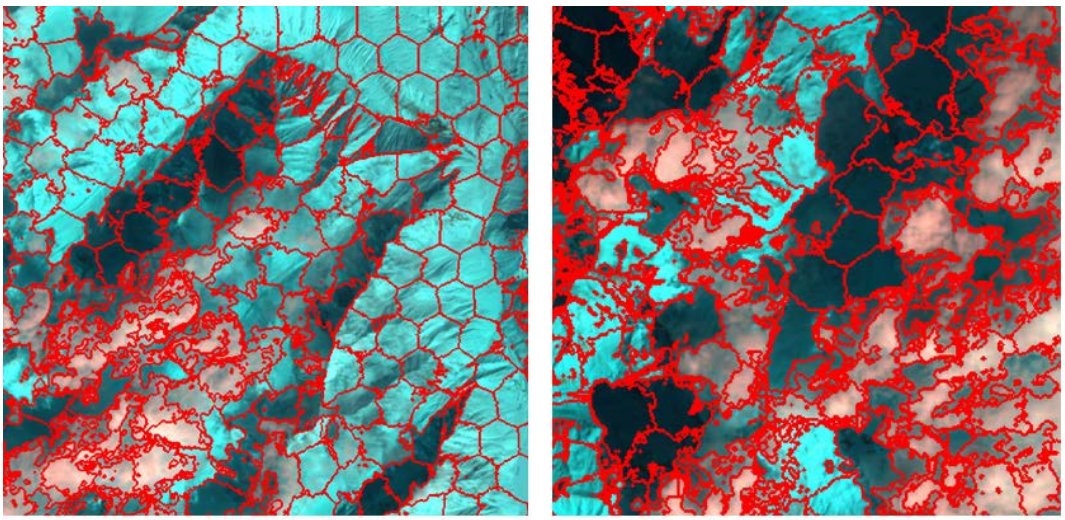

Figure 12: Example of the effectiveness of superpixels created with spectral angle mapper (beta version) to adhere to the edges of clouds, snowpack's, and shaded terrain. By classing at the superpixel scale instead of the pixel scale, only a fraction of the superpixels component pixels need to be positively identified for the class of interest to get accurate superpixel classification. This means many difficult to classify pixels (based on spectral signature) can be skipped and just assigned to the class of their neighboring pixels.

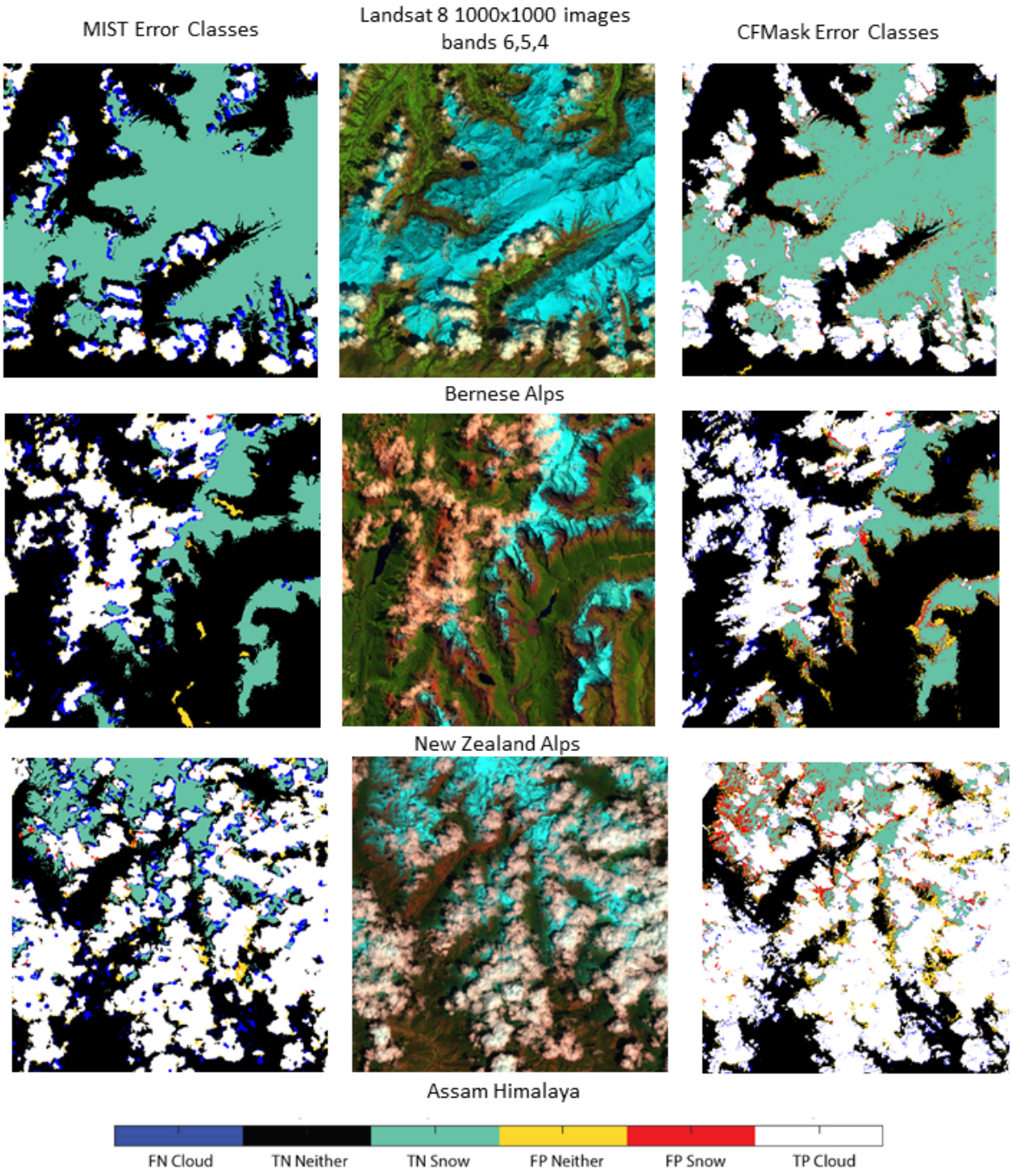

Figure 13. Examples of MIST cloud mask performance compared to CFMask over three areas with varying degrees of cloud cover. Pixels are colored by surface type and identified as false negative (FN), true negative (TN), false positive (FP) or true positive (TP) in the cloud mask. Top row Bernese Alps: Lötschental and Bietschhorn (3924 m), mostly snow with few clouds. Middle row New Zealand Alps: Climax Peak (2446 m) in the Pyke and Dart River Basins, clouds and snow about equal. Bottom row Assam Himalaya: below Chura Kang (6300 m) in Bhutan, cloud cover extensive but snow visible.

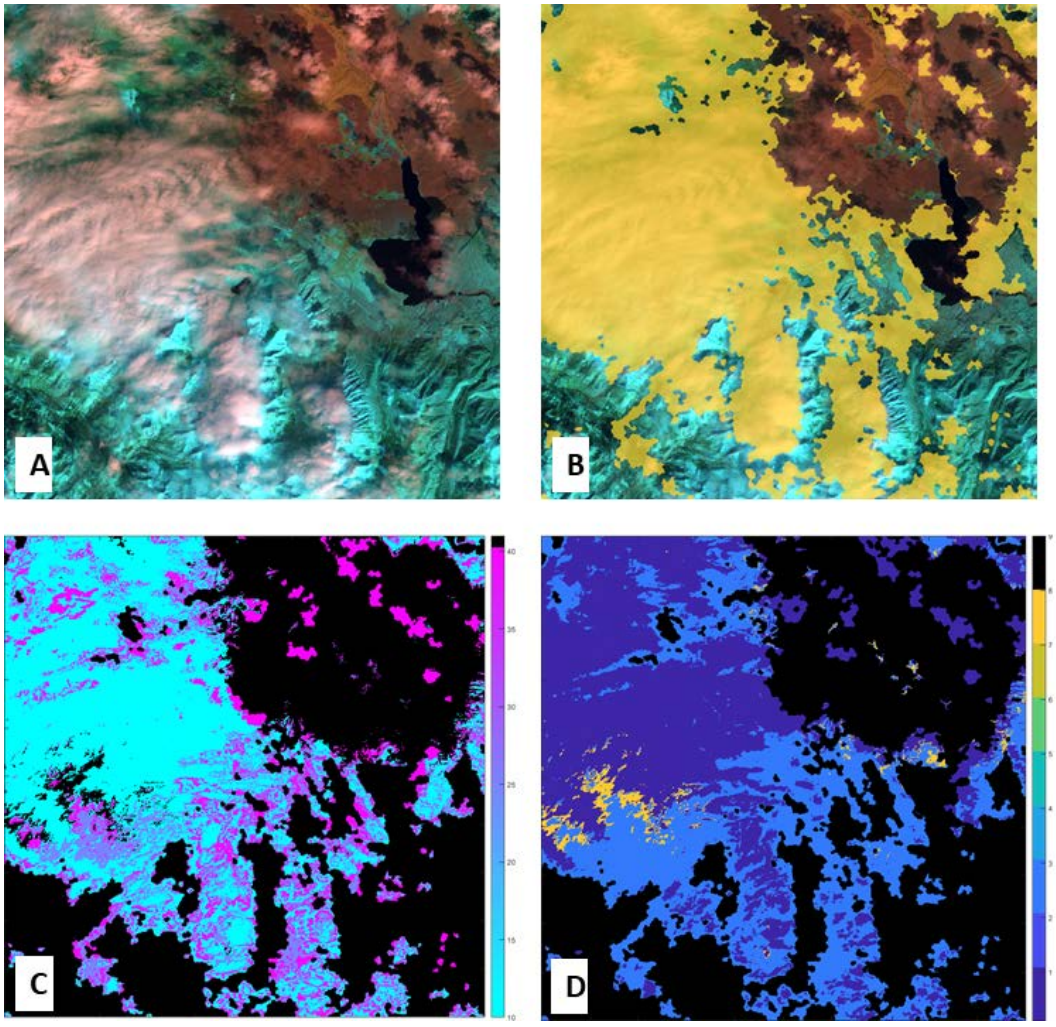

Figure 14: Example of output from the beta version of the new cloud mask. (A) false color RGB in Landsat 8 OLI bands 6, 5, and 3. (B) Yellow overlay of the cloud mask. Some cloud boundaries are not masked, the GDBSCAN cloud boundary module is not yet working. (C) Estimates of cloud top ice particle radius (um), black areas inside the cloud mask have particle radius estimates in the water cloud radius mask. (D) Estimates of cloud water equivalent (mm).

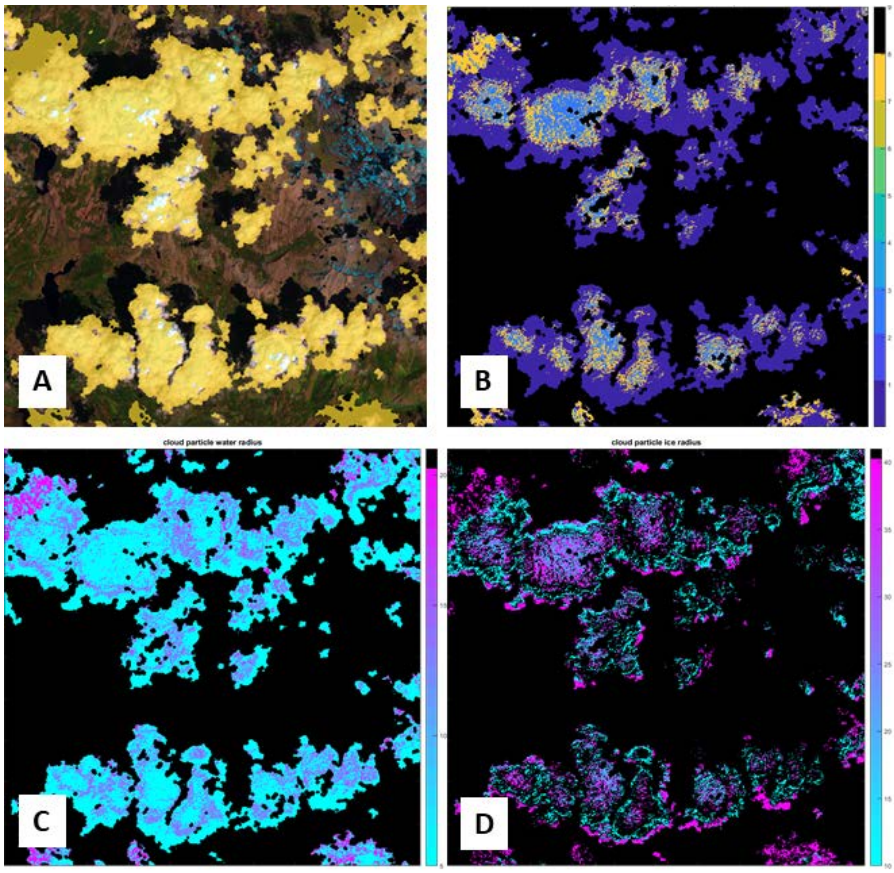

Figure 15: Example of output from the beta version of the new cloud mask. (A) Yellow overlay of the cloud mask on a false color RGB in Landsat 8 OLI bands 6, 5, and 3. Some cloud areas are not masked, but there is no false positive snow in the cloud mask. (B) Estimates of cloud water equivalent (mm). (C) Estimates of cloud water particle radius (um). (D)Estimates of cloud ice particle radius (um).

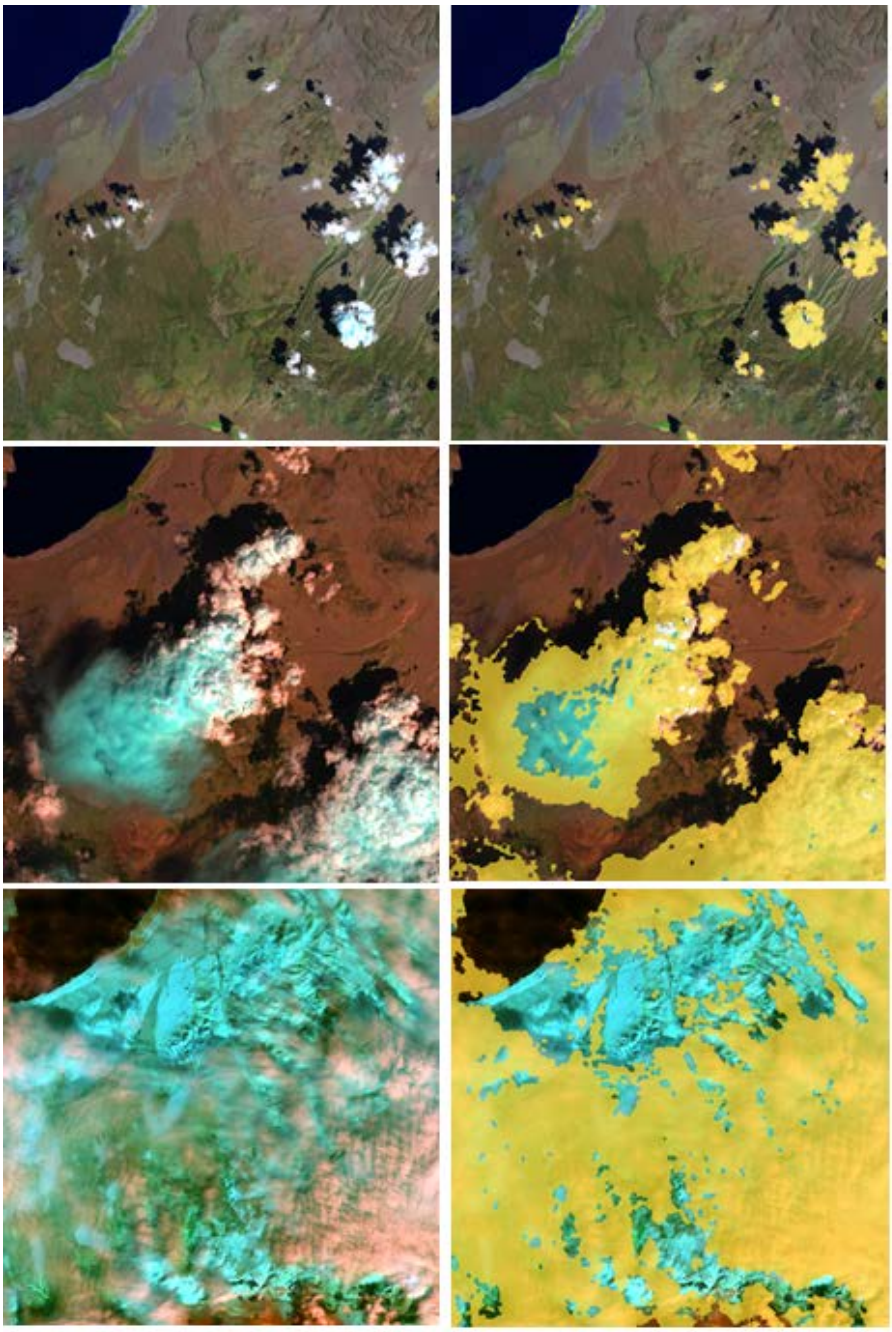

Figure 16: Example of cloud mask over the same location with varying amounts of snow and cloud cover. In the Top Panel, some clouds smaller than the size of the superpixels are missed. In the middle panel, some cloud shadows are misclassified as clouds and the center of a large cloud is missed. Bottom panel, cloud mask performs well with thin clouds over snow covered terrain, a difficult cloud type for per-pixel based cloud masks because of the transmission of some snow spectral properties through the thin clouds.

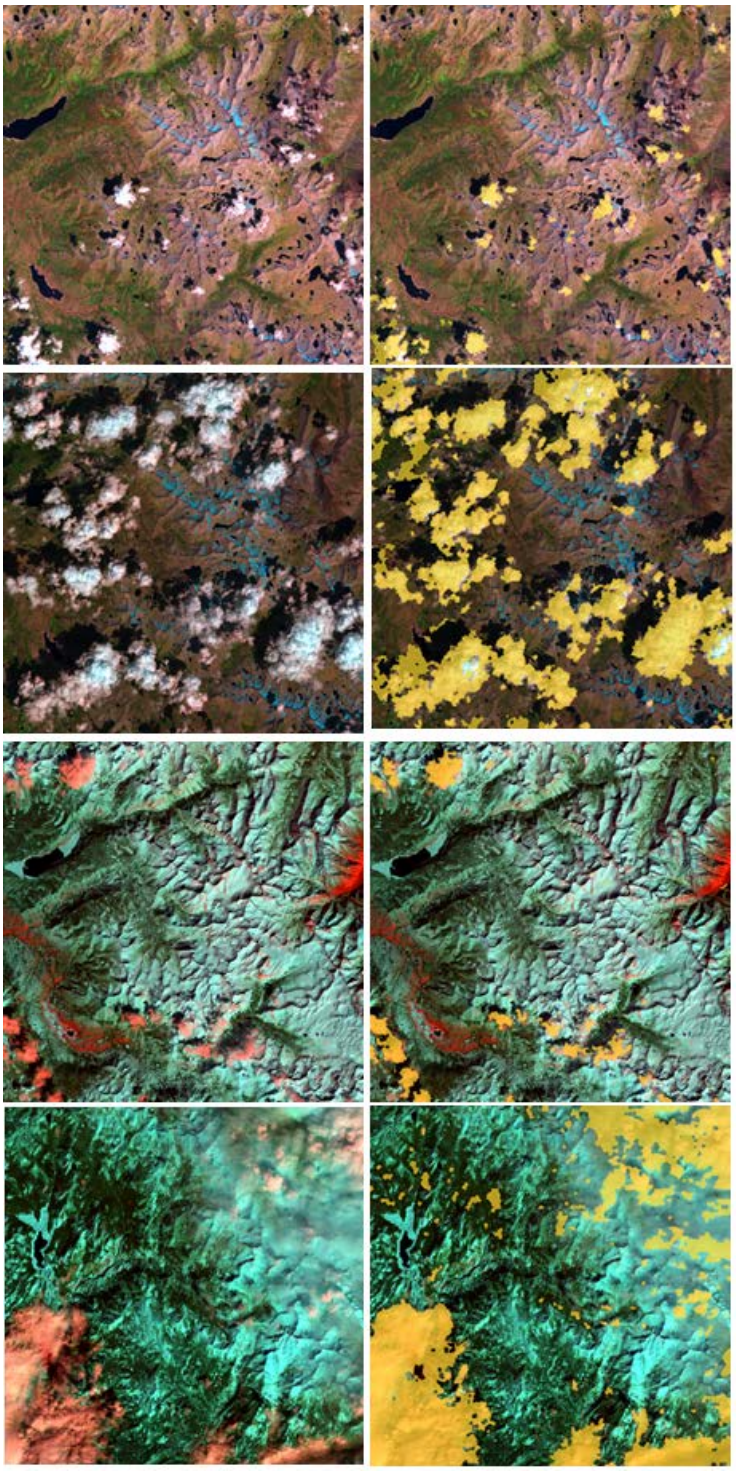

Figure 17: Example of cloud mask performance over varying levels of snow cover. Some cloud edges and centers are missed. Some areas of dark forest are currently confused with clouds. There are no pervasive issues of snow pixel inclusion in the cloud mask.

3 Impact of Water Supply Forecast Uncertainty on Water Management Decisions in California

3.1 Introduction

Mountain snowmelt is the primary water supply for over a billion people around the world, and recent estimates place the value of snow globally at over 1 trillion dollars annually (Sturm et al., 2017). Even with its economic significance, the harsh weather and terrain of the mountain environment constrain our ability to measure the state of the snowpack each year. In some mountain ranges, operational forecasts of snowmelt runoff do not exist. Even in well-instrumented mountains, forecasts are uncertain and sometimes have large errors. A pressing question applicable to both the gauged and ungauged regions of the world is, how useful are forecasts and how consequential are their uncertainty to water management? The traditional approach to answering these questions involves running hydrologic and decision optimization models, tuning parameters, and then observing modeled changes in operations. In theory, accurate forecasts have value. In practice, a forecast has operational value only if water managers respond to them. In this paper I empirically examine whether, and how, water managers respond to forecasts by implementing an empirical analysis on an historical record of water supply forecasts and their impact on water management decisions.

Water storage and delivery systems for surface water rely on a combination of predetermined allocations of water, operational forecasts, and flood risk management rules for guiding operations and determining water releases from reservoirs.

California's state water project (SWP), which provides water supply to more the 27 million people and 750,000 acres of farmland, is an example of how water management decisions are related to operational forecasts of water supply. The main driver for decision making at SWP, like many water projects, is delivering contracted water to their customers. Prior to the accumulation of winter precipitation, SWP receives requests for annual water allocations from the twenty-nine contractors that are supplied by the state project. In November, still prior to any significant annual precipitation, SWP determines the fraction of each contractors requested water they are allocated for the coming water year. These allocations are updated in public notices to SWP contractors, up to a few times a year, usually prior to April 1st. Measured accumulated precipitation and spring runoff forecasts from the California Department of Water Resources (CADWR) influence the decisions to increase or decrease the annual allocations of water. Similarly, the Central Valley Project, the U.S. Bureau of Reclamation's water supply project, issues initial water supply allocations for contractors between January and April of each year and bases their decisions in part on measured accumulation and CADWR runoff forecasts. For both of these large state and federal water projects, delivering allocated water supplies drives operations in the spring, summer, and fall.

CADWR forecasts are also used to set annual mandated environmental flows and are the foundation for the official water year type classifications in the Sacramento and San Joaquin rivers, the two largest rivers that drain the central valley of California. These index values are used to set strict minimum flows below reservoirs through the dry summer months (State Water Resources Control Board, 1978, 1999); managers

have access to the CADWR forecasts to prepare for required environmental releases dictated by the official water year type classification.

Additionally, surface reservoirs provide a valuable flood risk management service, reducing flooding in both the communities and irrigated lands below these mountainous watersheds. During the winter months, maximum stored volumes in each reservoir are independent of forecasted flows, allocated water supplies, or anticipated environmental releases required in the summer. Maximum stored volumes through the end of March are fixed to reduce the risk of flooding from rain during large winter storms. After the risk of flooding from winter rain subsides, flood pool storage is reduced and the focus of water managers shifts to delivering the contracted allocations of water, while maximizing hydroelectric generation where possible, and not failing to meet environmental flow requirements. Despite being constrained by these predetermined allocations of water and available winter storage, water managers still typically have the freedom to make subjective decisions about when to release water and how much to release.

Complicating California's water management is the link between its highly variable annual precipitation, limited surface water storage, and its economy. California, with a 2017 gross state product of $2.58 trillion (U.S. Bureau of Economic Analysis, 2018), over 39 million residents, and production of one third of all vegetables and two thirds of all fruits and nuts consumed in the United States, depends on a mountain snowpack for water supply. Either diverted from seasonal runoff or extracted from groundwater, Sierra Nevada snowmelt provides a continuous source of water for a state that experiences larger inter-annual fluctuations in precipitation than

any other State in America (Dettinger et al., 2011). Water management decisions matter every year as there is limited ability to capture surface water for intra-year storage in the reservoirs of California since the median total spring flows are similar in magnitude to the total storage capacity of the reservoirs in the Sierra Nevada (Rittger et al., 2016).

With dry summers and highly variable wet winters, the annual flows of winter rain and melting snow from the Sierra Nevada fill surface reservoirs and extend surface water availability into the dry summer months when demand rises, providing upwards of 70% of the state's annual water supply (California Natural Resources Agency, 2013). As this water drops down the mountain watersheds of the Sierra Nevada, energy is also harvested as hydroelectricity, contributing on average 15% of the state electrical generation ranging from 6% in severe drought years to over 20% of the power mix in wet years (California Energy Commission, 2018). California agriculture also depends on runoff and groundwater from the Sierra Nevada for growing valuable products with exports exceeding $20 billion in 2016 and gross revenues in excess of $50 billion in 2017 (U.S. Bureau of Economic Analysis, 2018). Urban water supply in the state of California is also tied to annual season flows and storage volumes in the reservoirs. Most urban water districts that supply water to California's 28 million urban residents rely on annual Sierra Nevada and Colorado River runoff alongside groundwater and local surface sources, and the diversity of their water sources buffers against the impacts of drought (Palazzo et al., 2017).

California has some of the best-instrumented rivers and monitored watersheds of the world's mountains because the Sierra Nevada are safe and accessible, water in

the state is such a valuable good with no substitutes, and there is limited storage space available. California has a century-long history of observing and measuring the snowpack in the Sierra Nevada (Church, 1914). Starting in 1930, the California Department of Water Resources began publishing an official bulletin called Bulletin 120 (B120) of forecasted snowmelt runoff for each basin of the Sierra Nevada, February through May of each year (California Cooperative Snow Survey, 1930-2019). These forecasts use multivariate linear regressions that predict annual spring runoff based on estimated year-to-date precipitation and expected precipitation (Huber & Robertson, 1982). Snow courses, linear transects of measured snow water equivalent, are the oldest form of field data used for the forecasts (Church, 1933). Eventually, the field data were expanded to include snow pillows, telemetered stations scatted around the Sierra Nevada in largely flat clearings at mid elevations. California has consistently invested in the effort needed to collect snow data for spring flood and water supply forecasts: in 2019 inflation adjusted US dollars, $220,000 per year in 1930, over $560,000 per year by 1954 (Strauss, 1954), to the latest published estimate in 2003 of $5.5 million per year (Roos, 2003). These forecasts are the baseline of available information water managers have for making informed decisions on storage and releases from reservoirs in preparation for the upcoming summer.

Despite the longstanding operational runoff forecasts, the State of California's Department of Water Resources predictions of the April-July flows, made on April 1 of each year, have a mean absolute error of 20%, one year in ten shows errors of 50% or more, and one year in fifty shows errors of 100% or more (Dozier, 2011). The field data for the forecasts are known to dramatically under-sample the snowpack,

especially at the lower elevations that are covered in snow during wet years and at the highest elevations, where snow can persist after it has melted off all the snow pillows (Rittger et al., 2016).

The accuracy of forecasts has been thoroughly documented in the western United States. Pagano et al. (2004) evaluated snowmelt runoff forecasts across the west, including one forecast record in California from the Walker River in the Eastern Sierra Nevada, and Harrison and Bales (2016) conducted a comprehensive analysis of the full history of California's Sierra Nevada runoff seasonal outlooks. Forecast skill increases as the season progresses. Areas like California with a wet winter followed by a dry melt season show forecast improvement through the season as the forecast shifts from a prediction challenge to a measurement challenge. In areas with wet springs and continued prediction challenges into the melt season, forecasts do not show improvement (Harrison & Bales, 2016; Pagano et al., 2004).

To date, no empirical analysis has been conducted to assess the historical record for how these operational forecasts impact water supply management. This paper is the first empirical test of whether these managers actually respond to forecasts. I hypothesize that volumes released from Sierra Nevada basins during the melt season are significantly associated with forecasts and their uncertainty of spring snowmelt runoff by testing two hypotheses:

A) The greater the forecast volume, the greater the release volume.

B) The greater the uncertainty in the forecast, the smaller the release volume

To test these hypotheses, a fixed-effect panel model is used to determine the association between California Department of Water Resources April 1st April-July runoff forecasts, their uncertainty, and April-July basin release volume.

3.2 Data

To test if water managers in California respond to forecasts, thirty-four years of water data from 14 high elevation basins were collected. The April through July water released from these basins were used as the measure of management decisions. Water releases from basins were modeled as a function of April through July runoff forecasts, their uncertainty, and available storage capacity the day before the forecast arrives. To control for basin-specific unobservable that do not change over time, a fixed-effects panel model (cross-sectional study through time) was used. Figure 19 displays a conceptual overview of the relationship between the data collected, how it is interpreted, and then used within the panel data model. Each component of the system in the conceptual model is explained in detail in the following subsections.

California was chosen as the study area because of the extensive public record of water data available to model the relationship between operational forecasts and water management decisions. To assist with the management of water supply, California has an expansive network of field measurements for monitoring how much water arrives as rain and snow in the mountains each winter, stream gauges for monitoring the flows of water, and measurements of the water volume stored in each reservoir. Not only has the state invested in measuring snow, rain, runoff, and reservoir storage, it has devoted significant resources to the care of statewide

hydrologic data and prioritized public access to these comprehensive hydrologic records. The California Data Exchange Center (CDEC: http://cdec.water.ca.gov/) houses these public data and exemplifies a database essential to the "Fourth Paradigm" of environmental informatics: *science driven by large volumes of data, integration of disparate datasets, and their associated computational challenges* (Frew & Dozier, 2012; Hey et al., 2009). CDEC uses a three-character alphanumeric coding scheme to identify stations and a numerical scheme to key sensors that collect information at different stations. Contemporaneous records were aggregated for this study.

Figure 18 gives an overview of the study location. The high elevation watersheds are located along the interior of the state with the highest, snow dominated watersheds to the south, and the lower elevation watersheds more influenced by rain to the north. There are a mix geologies and hydrologic characteristics present among the study sites as well. Heavily granitic watersheds, mainly in the southern sierra, watersheds with a mix of metamorphic and granitic base rock in the central sierra, and the northern Sacramento dominated by volcanic soils from Mt. Shasta. Among the 14 watersheds within this study is a mix of water management structure, hydrology and climatology.

3.2.1 Monthly Full Natural Flow (FNF)

Full Natural Flow (FNF) is a calculated value of the unimpaired flow in a river basin that adjusts the measured flow for six ways that human actions and natural processes alter that unimpaired flow: changes in reservoir storage, reservoir evaporation, exports to other basins, imports from other basins, and finally diversions

minus return flows from consumptive uses within the basin. I use the volumetric measurement over each month. California Department of Water Resources, water operators on each river, and the US Army Core of Engineers are responsible for the calculations. Monthly FNF data from the same locations as the runoff forecast for each of the 14 Sierra Nevada basins in Figure 18 are used to calculate the forecast error of the CADWR spring runoff forecasts. These stations are the same stations for which the seasonal runoff forecasts are produced.

3.2.2 California Department of Water Resources April 1st Forecast and Uncertainty

The April through July runoff forecasts are contained within the State of California "Bulletin 120" (B120) annual report. The B120 forecasts of the Monthly Full Natural Flow (FNF) for these 14 Sierra Nevada Basins are the longest running forecasts in California of water resources from the Sierra Nevada, in some watersheds dating back to the 1930s. The CADWR forecast of spring runoff used in this study is the B120 seasonal bulk volume forecast which includes a confidence range (90%-10% exceedance probabilities). These forecasts are issued monthly February thru May and are based on regression analyses from measured snow water equivalent, rain, and streamflow. The April 1st April-July31st 90%, 50%, and 10% exceedance probability FNF forecasts are used in this analysis. The April 1st forecast is used because it is the last forecast available before the start of the forecasted time period and represents the knowledge available to water managers at the start of the spring runoff season. The 50% probability forecast is used as the forecast value and the uncertainty in the

forecast is calculated as the 10% exceedance forecast minus the 90% exceedance forecast. This measure of uncertainty represents the level of uncertainty available to water managers at the time of the forecast and is variable from year to year. B120 forecasts were obtained directly from personnel at the California Department of Water Resources for the years 1930-2011. The 1985-2011 forecasts were used from this database. The 1999-2018 forecasts were extracted from the available online records of B120 reports. The overlapping region of 1999-2011 between the two sources of B120 data were used to QA/QC the two datasets.

3.2.3 Forecast Error

To derive forecast error, I compared the 50% forecast to the full natural flow dataset with the following equation:

$$Forecast\ Error\ = \frac{Forecast - Full\ Natural\ Flow}{Full\ Natural\ Flow} \quad (1)$$

Negative error indicates an under forecast, there was more flow than expected; positive errors indicates an over forecast, less flow than expected. Negative error magnitude has a limit of -100%. In the positive direction, there is no limit on the magnitude of the forecast error.

3.2.4 Monthly Basin Releases

Basin releases—defined as the measured monthly flow below the terminal reservoir for the 14 Sierra Nevada basins in Figure 18—is the volume of water that was allowed by water managers to leave each basin and is the best available integrated value that reflects the cumulative decisions of water managers in each basin each month. Because the overall storage capacity of the basins is about equal to the mean

annual flow, most water that flows out of the basins results from decisions. Apart from environmental flows, water managers have freedom to make subjective decisions about when to release water and how much to release. Table 10 shows the variability in the difference between FNF and basin releases and gives an overview of the datasets used in the panel analysis and the variability between the 14 basins used in this study.

3.2.5 March 31st Available Reservoir Storage.

Monthly end-of-month reservoir storage data are available for 74 reservoirs within the 14 basins in this study area. As shown in Figure 19 the individual reservoir data were summed to the basin level to calculate available reservoir storage at the same resolution as the monthly flow data and the seasonal runoff forecasts. All reservoirs in the basin are treated as a single large reservoir. Available storage space in each basin on March 31st is calculated by taking the total basin storage capacity and subtracting measured stored volumes. Total basin storage capacity is the summed volume of all reservoir storage capacity, not adjusting for dead pool, in each basin. These values represent the starting condition for water managers when they receive the April 1st forecast and begin active management of the spring runoff. Calculated available storage space is used instead of measured stored volume, as available space is a direct representation of the conditions managers must work with and avoids having to calculate lost storage capacity due to sedimentation or worry about the specifics of each reservoirs usable pool. The list of all reservoirs used in this analysis with information on their capacity, which basin they are in, their elevation, and which organization operates them are in Table 13. The monthly reservoir storage data was

the limiting factor for extending the study to before 1985; too many reservoirs had no available data in CDEC prior to 1985.

3.3 Statistical Analysis

To examine the associations between the volume of water released each melt season from the 14 western Sierra Nevada watershed with the predictors described above, a panel data model with fixed individual effects and standard errors robust to heteroskedasticity and adjusted for 14 clusters was used (Álvarez et al., 2017).

For this study, the unit of analysis is each of the fourteen basins in the western Sierra Nevada at an annual time step over the years 1985-2018. The starting year of 1985 was chosen because it is the first year that data were available for all parameters in the panel data model. Because a few data points necessary for a complete record are missing from CDEC, the model was run as an unbalanced panel with n = 416 and 398 degrees of freedom. A fixed effects panel model is used to control for unobserved heterogeneity in each basin. The form of the panel model used is:

$$R_{it} = \beta_0 F_{it} + \beta_1 S_{it} + \beta_2 U_{it} + \gamma_1 F_{it} U_{it} + \mu_i + \epsilon_{it} \quad (2)$$

Where R_{it} is the release of water from basin *i* in year *t*, F_{it} is the forecasted flow, $S_{it}$is the available storage capacity the day before the forecast arrives, $U_{it}$is the uncertainty in the forecast, and $F_{it}U_{it}$ interacts the forecast with its uncertainty. The fixed effect μ_i accounts for basin-specific unobservables that do not change over time and the error term is ϵ_{it}. The error term is assumed:

$$E[\epsilon_{ti}] = 0 \; with \; i.i.d\left(0, \theta_\mu^2\right) \quad (3)$$

Error is assumed to have normal distribution with a mean of zero and no autocorrelation. This setup enables modeling the relationship between water supply management and spring runoff forecasts by evaluating the associations of forecasts with basin releases and forecast uncertainty.

3.4 Results

Table 11 summarizes the results of the Panel data model shown in Equation *(2)*. The fixed effects panel data model tested the association of basin releases and water supply forecasts. The model coefficients estimate the marginal change in the outcome variable, basin releases, by changing one unit of the predictors, controlling for the heterogeneity between groups. Beta values represent the average effect within units. By analyzing associations at the basin scale, the preferred model was able to predict 70% of the variability in the outcome of April-July basin releases (adjusted $R^2 = 0.70$, F-Test, $p \ll 0.001$). The April 1st forecast, April 1st forecast range, available storage capacity, and the interaction between forecasted volume and uncertainty were all statistically significant predictors ($p \ll 0.005$) of basin releases. Increased forecast uncertainty and increases in available storage were both negatively associated with April-July basin release volume, whereas forecast volume and the interaction between forecast uncertainty and forecast volume were both positively associated with release volume. The interaction term, γ_1, is positive which means that as forecast volume increases, the effect of uncertainty decreases and, in the model, the effect of uncertainty disappears for a forecast above about 3.5 km^3. On average, when more than 3.5 km^3 is forecasted, there is no effect of uncertainty in the model. Results

suggest that all else being equal, basins with larger forecast uncertainty on April 1st release less water.

In preparing the data for the panel data model, the error distribution for February, March, April, and May forecasts was also analyzed. Figure 20 shows the relative error distribution of all historical forecasts for all basins in the study. On the negative end, errors can never be worse than –100%. On the positive end, one forecast in ten shows errors of +50% or more, and one forecast in fifty shows errors of +100% or more.

February and March are significantly worse than the April forecast, and results matched Harrison and Bales 2016 analysis (Harrison & Bales, 2016). May 1st April-July forecast has higher performance, based on the Nash-Sutclif coefficient for each basin, than the April forecast as seen in Figure 22. This figure also shows that the published April-July forecasts in April and May have different error distributions. But, when the May forecast is adjusted by removing measured April FNF from the forecasted flow volume, the error distribution is actually the same for April and May forecasts, with no improvement in forecast skill from April 1st to May 1st. A t-test failed to reject the null hypothesis that there is a difference in the mean Nash-Sutcliff coefficient between the April forecast of April-July runoff and the May forecast of May-July runoff. The March and February forecasts are both significantly worse than the April 1st forecast.

3.5 Discussion

Reservoir operators appear to be risk adverse to the possibility of less spring flow than is forecasted. Increased uncertainty in official April-July runoff forecasts

significantly reduces the volume of water western Sierra Nevada basins release April-July. When a forecast is more precise for a basin in a given year, reservoir operators have higher confidence that additional inflow will refill reservoirs, and the model confirms that more water is released earlier in the year, which may enable additional beneficial uses of the earlier released water possibly not available had it been released later in the summer. A tractable goal for driving forecast improvement that may impact reservoir operations, yet not require reducing error in forecasts, is to aim for reducing the uncertainty in runoff forecasts.

The model also showed that risk aversion decreases as the forecasted flows increase. This effect is seen in the positive sign of the coefficient on the interaction term between forecast volume and forecast uncertainty. As forecast volume increases, the effect of uncertainty decreases, most likely due to the ratio between the volumes of water forecasted to the available storage space in the basin's reservoirs. If there is a forecast with a large amount of uncertainty, a wide spread between the 10% and 90% exceedance probabilities for this study, but for ample water relative to the available storage space in the basin's reservoirs, then the model confirms common knowledge that the risk aversion effect decreases and eventually disappears when there is far more water expected than room to store it. Because of the high variability in hydro climate in California, this is an important but infrequent situation as many years do not experience flow volumes large enough to reduce the effect of uncertainty.

Prior analyses on the B120 forecasts have focused on the forecast error on its own and not directly connected forecast performance to reservoir operations (Dozier, 2011; Harrison & Bales, 2016; Pagano et al., 2004). Reservoir operations models predict

forecasts to return the most value to water managers when reservoir storage capacity is between 25-100% of the mean annual flow volume of the river and that significantly larger reservoir operations will not respond to improved forecasts (Barnard, 1989; Odegard et al., 2019). Analysis from this paper shows that basin releases across 14 Sierra Nevada basins, which have reservoir storage capacity similar to annual flow volumes, are sensitive to runoff forecasts, which agrees with Barnard and Odegards findings; the ratios between basin storage capacity and mean annual flow in this study are similar to these previously reported ratios that are sensitive to forecast quality. Prior similar work in the Sierra Nevada focused on hydropower revenues used models to see how better forecasts may lead to increased profits, with values of $1 million calculated for the Yuba-Bear system (Rheinheimer et al., 2016). Those results explicit to hydropower profits on a single basin of the Sierra Nevada, showed that uncertainty reduction increased forecast value, in part because of the volume of available storage capacity relative to mean inflows. This study also extends these prior findings by showing value in uncertainty reduction beyond Yuba-Bear hydropower to all western Sierra Nevada basins, which have diverse operations beyond just hydropower, ranging from water supply for urban and agricultural uses, to flood control, and environmental flows.

Recent advances in forecasting that incorporate distributed hydrologic sensor networks (Zhang et al., 2017), remotely sensed data (Rittger et al., 2016), distributed hydrologic models or adaptive flood-control measures that can be set based on high-resolution forecasts (Jasperse et al., 2017), have all been shown to improve forecasts. These new methods are not yet operational across the state and do not usually include

an analysis of the connection between improved forecasts and their impact to actual water supply operations. This analysis shows the relationship between existing operational forecasts and reservoir operations and that there is room for improvement above existing operational forecasts. If certain aspects of the forecast are better, specifically more certainty in upcoming flows, implementing some of the new research into operational forecast could lead to changes in water operations.

3.5.1 Water Operations and their Relationship to Hydrology

To optimally manage the water supplies in the Sierra Nevada, water resource managers must have knowledge about the precipitation accumulation over the winter, its distribution between rain and snow, the quantity of water within the snow (i.e., the snow water equivalent), when and how quickly snow will melt, and how much of that water will flow into the distribution system. The longer the decision maker waits for improved information, the greater the opportunity cost for not acting, especially if the desired actions require mobilization of resources or coordination with institutions or individuals. This tradeoff between the value of an early forecast versus the value of a later but more robust forecast has been explored under a general class within optimal stopping theory known as the "commitment" problem under forecast uncertainty: decisions on allocations of water must be made before the runoff has occurred (Krzysztofowicz, 1986).

California's water infrastructure, water rights, and institutional framework for managing water pre-date many of the recent advances in forecasting and hydrology that could improve forecasts. Since large decisions on water allocation require

significant coordination between parties and the process of issuing water allocations or revising allocations takes time, historically the benefit of extra time afforded by setting water allocations early in the winter, far outweighed the cost of waiting into the spring to make major decisions, when the probability of a poor forecast with operational methods was still very high. Current policy, regulation, and precedent in California is such that the timing of many major decisions, like the early winter annual allocations for SWP and CVP contractors, are fixed and still precede the availability of a good operational or research forecast.

The risk averse behavior seen in reservoir operators may be a result of reservoir operators striving to meet promised allocations of water that are set in contracts prior to a full understanding of the volume and timing for water to come as runoff. Since many commitments will have been made prior to knowing the annual water supply, containing the uncertainty is an actionable improvement that may still have an effect on improving the timing of releases and avoids the much more challenging task of trying to change the timing of decisions made for major water allocations in California to a later date.

Some error and uncertainty in forecasting is inevitable. The uncertainty can be divided into two broad categories: uncertainty in measurement of year to date precipitation and uncertainty in expected precipitation volume. In early winter, when many operational decisions are being made, before seasonal snow has fallen, spring runoff forecasts begin as a prediction of the volume of water that might accumulate in the coming year. Operational runoff forecasts in California rely on historical medians for predicting upcoming seasonal precipitation (Roos, 2003). As the season progresses,

the forecast fundamentally changes to an estimation of likely runoff based on in situ (Cox et al., 1978) combined with expected precipitation to come.

Figure 21 shows the progression of uncertainty through the winter season in an official use of the spring runoff forecasts for guiding release volumes. The Sacramento Valley and San Joaquin Valley Water Supply Indices, which set legally required environmental flows, are calculated and set using the May B120 forecast. Table 10 shows which B120 forecasts are aggregated to calculate these two water supply indices. Errors in B120 forecasts directly impact the water supply forecasts. Since 1995, the B120 forecasts have been implemented in the following way to calculate the water supply indices: On the first of each month, from Dec to May, official forecasts and confidence bounds are given for the water year classifications with the May 1st forecast 50% exceedance used to assign the Sacramento Valley Water Year Type and the 75% exceedance used to assign the San Joaquin water year type. Official Water Supply Index forecasts each month include a 90% and 10% exceedance probability confidence range in the expected water year type. The forecast record includes the final calculated observed water year type, which can be different from the official May 1st water year type. Figure 21 shows the distribution of error in the forecast of water year type as the winter progresses for the full record of water year type forecasts from CADWR. Zero on the y axis represents the final, calculated water year type based on observed runoff at the end of the water year. The ordinal water year type classifications are measured in distance away from the final observed water year type. The official categorical scale for water year type is 5 units wide (critical, dry, below normal, above normal, wet). Red boxplots represent the distribution of the 10% exceedance probability category in

distance away from final observed water year type, the blue boxplots represent the distribution of the 90% exceedance probability in distance away from final observed water year type, for the full historical record of water year type forecast. The May forecast is the official water year classification set by the state and used to guide summer reservoir operations, which can still be different than the measured water year type that is calculated at the end of the water year. In December and January, the predicted range the water year type will fall within commonly extends across all water year types. By April, there is still a range of ±2 water year types. When the official water year type is set in May, the actual water year type can be ±1 water year type away.

The progression of the uncertainty around the office Water Supply Indices for the San Joaquin and Sacramento Rivers, highlight the information environment within which some of these large decisions are currently made. The Sacramento and San Joaquin water supply indices are stipulated (State Water Resources Control Board, 1999) as the official index for determining environmental releases in these two rivers during the water year. These indices are the only major seasonal runoff forecast issued by the state staring in December, and their quality provides a good look into the operational forecast information quality at the time of year when major water allocations decisions are made early in the winter. These water supply index forecasts have a wide uncertainty range early in the year which can sometime span the full range of possible water year types. The ability to predict water year type in December using the current operational methods is poor yet this is the time when major SWP water allocation have been made, and many CVP allocations are about to be made.

The impact of forecast error on water year type is clear from Figure 21. The B120 forecast errors impact preparation for required minimum releases and mandated minimum releases through the summer. Volume and timing of releases directly affects both environmental flows and hydropower production options. As minimum flow requirements are set by water year type, in years where more uncertainty exists and the May 1st water year type is unknown, less water may be released. Water managers may be holding back water to hedge against the water year ending up direr than predicted.

The May 1st B120 forecast sets important flow requirements for the summer months and as discovered here, may be less informative of future flows than currently assumed in the literature. The standard comparison, when these sequential forecasts are compared, is to compare skill measures of the April-July forecast made each month. Using this methodology May 1st forecasts, which are used to set Water Year Type and occur during the melt season, are of significantly higher quality than the April 1st forecast (Harrison & Bales, 2016; Pagano et al., 2004). Unfortunately, 20-30% of April-July runoff occurs in April, and even though May 1 forecasts are an improvement over April 1, they "predict" a significant fraction of streamflow that has already occurred. A close look at Figure 22 shows that May 1st forecast improvement is not significantly better than April 1st at predicting the volume of water still to come. The errors in May predictions of May-July runoff and April predictions of April-July runoff come from the same statistical distribution. The traditional method of comparing various forecasts of April-July runoff overestimates the increase in forecast performance in the May B120 forecast.

3.5.2 Operations Forecast Assumptions and Available Improvements

Snowfall and precipitation occur irregularly over the winter season and variations in the timing of accumulation lead to larger B210 forecast uncertainty due to the fact that all yet to fall precipitation is modeled as median values in B120 and Water Supply Index (WSI) forecasts. California derives a majority of its annual precipitation in 6-10 large storms. The timing and spatial distribution of precipitation phase of these events can cause median assumptions to misguide projections of the effect of additional precipitation on runoff volume and timing.

Seasonal prediction ability exists that can inform water managers well ahead of actual precipitation at the basic level if median assumptions will be less appropriate for the given year's climatology. Statistical models based on climate indicators, such as the El Niño-Southern Oscillation (ENSO), can forecast the probability of the seasonal precipitation tercile – wet, dry, or normal - before the season begins (Piechota & Dracup, 1996) with recent improvements enabling discrimination among the types and magnitudes of ENSO events improving these seasonal climatology predictions (Williams & Patricola, 2018). Once the season starts, climate indicators such as the state of the Madden-Julian Oscillation possibly provide statistical predictions six to ten weeks ahead (Guan et al., 2012). These statistical, seasonal, and 8-10-week forecasts are not currently factored into determining the likelihood of median conditions based on the existing precipitation volume of the year so far. Median assumptions are easy to implement and use to guide regulations, but the potential increase in forecast uncertainty can reduce the benefit of using the water. A partition of the forecast into parts—separate forecasts of expected flows from precipitation on the ground and of

predicted additional precipitation in the season to come—would be a simple method to better categorize the type of uncertainty, provide greater clarity under existing guidelines, and could help narrow the uncertainty of the forecast of future flows.

Another opportunity for positive impact from improving existing use of forecasts is through forecast informed reservoir operations. During the winter months, maximum volumes allowed to be stored in each reservoir are independent of the forecasted flows and set to reduce the risk of flooding from big winter storms. The draft plans for forecast informed operations of Lake Mendocino showed that existing fixed rule curves can force winter releases that may otherwise help fill reservoirs in dry years (Jasperse et al., 2017). If reservoirs can be left filled above their flood rule curve when there is high confidence of dry conditions during the winter flood risk management stage of reservoir operations, they will enter the spring melt season with less available storage space for water, which the panel model showed was a significant association with increased releases during the April-July time period.

In addition to the existing in-situ observations used to forecast water supply, satellites reliably map snow cover extent and albedo in the Sierra Nevada and mountain ranges worldwide, data which can inform runoff forecasts but are not incorporated in the B120 forecast. Rittger et al. (2016) showed that remote sensing imagery can show existence of yet-to-melt snow when all sensors are snow-free and that the relationship between fractional snow-covered area and streamflow becomes reliable during the period when B120 forecasts are still error prone. Quantitative tools currently exist that may provide significant improvement to estimates in April and May. These range from large operations like the NASA Airborne Snow Observatory

(Painter et al., 2016) to simpler options such as sensor interpolation constrained by satellite-mapped snow cover (Dozier et al., 2016; Fassnacht et al., 2003). Even just interpolating snow water equivalent between the existing sensor networks constrained by snow covered area may provide significant improvement over operational forecasts. The sensor network is also a legacy network that provides statistical power from its long historical record for which it has been used. But with climate change, as the past becomes less representative of the future, the power of these forecast may be reduced. Current work on improved sensor networks that accurately measure the spatial and temporal variability of perception, snow depth, wind, temperature and humidity (Zhang et al., 2017) could help improve the operational network and contribute to an effort towards making our forecasts robust to changing trends from climate change.

Additionally, with climate change and more precipitation coming as rain instead of snow, less certain forecast from earlier in the season may become more important as the spring flows shift to earlier dates. Currently, our ability to measure a snowpack is a more robust integrated measure of precipitation than trying to measure snow accumulation or rainfall. As the fraction of annual water that comes as rain increases, this is a concerning issues that impacts our ability to forecast and manage the surface water supply.

3.5.3 New Opportunities for Improved Forecasts

The observed reductions in April-July basin releases under higher forecast uncertainty is contrasted with significant economic opportunities for water that water

managers maybe be foregoing during the melt season. Increased allocations to irrigation and urban water districts, electricity generation, instream environmental benefits, and groundwater recharge opportunities lost are all potential opportunity costs with less water released April-July than would be with less uncertainty.

One new major opportunity for additional value from improved forecasts is California's Sustainable Groundwater Management Act (SGMA) and a general trend in California toward a more integrated approach to ground and surface water management. An effective method for reducing the chronic overdraft of California's ground water resources is opportunistic groundwater recharge using known excess surface water flows from the spring melt season. California Department of Water Resources has implemented these ideas in the Flood-Managed Aquifer Recharge (Flood-MAR) program, a volunteer program that encourages opportunistic use of flood waters for beneficial water management (CADWR, 2018). Some Flood-MAR opportunities may exist only early in the melt season and benefit from increased basin releases in these earlier months of the runoff season. With less forecast uncertainty, this model suggests that more water could be released from basins April-July, and the water could be diverted downstream to ground water recharge during these months.

The timing of available flows has an impact on the type of land available to receive opportunistic water for ground water recharge. Agricultural land is a possible source of land for Flood-MAR and the timing of available water has a large impact on whether or not the land is available or how effective the recharge may be. Earlier flows could flood fields when evapotranspiration is lower, before fields are planted, or enable farmers to plant crops more tolerant to or that benefit from flooding. Water

temperature of flood waters could be lower for shallow flooding of fields earlier in the year as well. Earlier flooding could enable land to be used for Flood-MAR that may be unavailable later in the growing season when planted crops are approaching harvest and flooding is not compatible with crop growth or harvesting requirements when farm machines and labor need drier soils to access fields.

Conjunctive land use to recharge groundwater requires coordination with landowners and time to execute. Higher certainty in excess flows earlier in the year enables time for that process to occur prior to the availability of the water. Beneficial uses for water earlier in the season could also increase reasons to send water down riparian corridors during the spring melt season, a time when ephemeral ecosystems historical received water prior to reservoir construction. The opportunity to increase late winter and early spring flows provides environmental benefits in the rivers as water is transported through them to recharge groundwater.

Having higher confidence in expected inflows in the high elevation reservoirs increases the ability to spread the diversion of water for possible Flood-MAR projects over a wider period of time. Trying to quickly divert water is difficult as the infrastructure of canals and diversion ditches is more effective handling a slower, longer duration flow through the small channels as opposed to being overwhelmed by a sudden need to divert a large volume. The same opportunities for groundwater recharge may not exist for the water if it is released as spills or unexpected releases during the melt season, as required planning and coordination cannot instantaneously take advantage of unexpected flows. Willingness of high elevation storage facilities to

increase spring releases depends on reducing their uncertainty in total spring runoff volume.

3.5.4 Strengths and Limitations of the Analysis

In this paper I did not look at the actions of the water managers from August through March, while the results showed that there was less water released during the spring melt season when forecast uncertainty increased, I did not identify if these releases are just delayed into the future summer months or if they are released prior to April. Additionally, the unit of analysis could have be broken down further to the individual operating agencies within the Sierra Nevada or the individual reservoirs themselves to test at a finer resolution the impact on these individuals and how responses might vary given the different management priorities of these more focused subsets. The April-July timeframe was chosen because a large fraction of annual runoff occurs during this time period, forecasts are available at the start of the time period for the expected flows, and reservoir operators are actively controlling the volumes and timings of basin water releases. The basin level was chosen because it represents the overall picture well, and most of the data is at that scale. The significant findings at the coarse resolution give merit to a more detailed look at operators or individual reservoirs. The surface water supply system also has constraints on operations that are independent from river flow volumes; flood control, environmental, and water right constraints on operations dictate specific actions each year that may reduce the explanatory power of a forecast on actual basin releases. While the hydrologic

conditions each year are highly significant in how much water is released from each basin, none are perfectly correlated with releases.

These significant findings in California bolster the case to look at similar systems in other states or parts of the world where the timing of decisions and operational ecosystem for managing water are different and may confirm this paper's findings or lead to different outcomes. For example, in High Mountain Asia, where no operational observations or forecasts exist, uncertainty is much higher and understanding how this impacts water allocations and operations may inform pathways to improving water supply management in more austere regions.

3.6 Conclusion

Analysis from this paper shows basin releases across 14 Sierra Nevada basins are sensitive to runoff forecasts. Reservoir operators in the Sierra Nevada of California are risk adverse to over predictions of spring runoff. Increased forecast uncertainty and increases in available storage were both negatively associated with April-July basin release volume, whereas forecast volume and the interaction between forecast uncertainty and forecast volume were both positively associated with release volume. Results suggest that all else being equal, basins with larger forecast uncertainty on April 1st release less water April-July. The model also showed that risk aversion decreases as the forecasted flows increase.

Seasonal prediction ability exists that can better inform water managers than the current operation forecasts. Quantitative tools that harness distributed sensor networks, hydrologic models, and remote sensing data currently exist that may

provide significant improvement to runoff estimates in April and May. If certain aspects of the forecast are better, specifically more certainty in upcoming flows, implementing new research into operational forecast could lead to changes in water operations. A tractable goal for driving forecast improvement that may impact reservoir operations, yet not require reducing error in forecasts, is to aim for reducing the uncertainty in runoff forecasts.

Accurate forecasts enable planning for optimal use of runoff, a variable and important resource. Increasing spring flows would enable alternative uses of water during the spring runoff period and help with the State of California's goals to better integrate surface and groundwater management.

An important note to water managers and regulators across the state is that there is a higher risk than expected - based on the current use, analysis, and dissemination of the May 1st forecast - for it to have a significant error in expected May-July inflow. There is no improvement in B120 forecast skill from April 1st to May 1st. A more detailed presentation of the spring runoff forecasts that communicates estimates of measured precipitation and measurement uncertainty separately from expected precipitation and forecast uncertainty could better inform and guide water operations in the state.

This study highlights how improvements to operational spring runoff forecasts could directly translate into changes in water management in California and open new opportunities for using surface water.

3.7 Tables

Table 10: Overview of the surface water system on the west side of the Sierra Nevada used in the panel analysis

River System	Basin	Bulletin 120 Forecast Point	Latitude	Longitude	Basin Storage Capacity (Km^3)	Reservoirs	Drainage area above reservoirs (Km^2)	Precipitation Stations†	Snow Courses†	Snow Pillows†	April-July Release as % of AJ FNF median (min,max)
Sacramento	Sacramento	Sacramento above Bend*	-122.186	40.289	6.1	8	17262	8	12	10	130 (91, 220)
	Feather	Feather at Oroville*	-121.547	39.522	6.7	11	9342	8	22	10	70 (24, 150)
	Yuba	Yuba at Smartville*	-121.274	39.235	1.85	8	2849	10	20	5	67 (28, 152)
	American	American at Folsom*	-121.183	38.683	2.17	12	4882	10	28	17	87 (49, 235)
San Joaquin	Cosumnes	Cosumnes at Michigan Bar	-121.044	38.500	0.06	1	122	7	8	0	92 (66, 101)
	Mokelumne	Mokelumne at Pardee	-120.719	38.313	1.07	4	1603	8	19	10	71 (43, 137)
	Stanislaus	Stanislaus at Melones	-120.637	37.852	3.5	8	2331	9	17	10	38 (12, 89)
	Tuolumne	Tuolumne at Don Pedro**	-120.441	37.666	3.44	6	3994	9	21	10	49 (7, 102)
	Merced	Merced at Exchequer**	-120.331	37.522	1.28	2	2686	7	13	11	79 (30, 176)
	San Joaquin	San Joaquin at Friant**	-119.724	36.984	1.42	8	4338	8	21	16	77 (4, 100)
Tulare	Kings	Kings at Pine Flat	-119.335	36.831	1.53	3	4002	8	25	14	106 (74, 156)
	Kaweah	Kaweah at Terminous	-119.003	36.412	0.23	1	1453	10	16	8	100 (47, 151)
	Tule	Tule at Success	-118.922	36.061	0.1	1	1018	10	7	3	101 (56, 361)
	Kern	Kern at Isabella	-118.484	35.639	0.7	1	5372	8	24	12	100 (60, 139)

* Used to forecast and calculation of Sacramento Valley 40-30-30 Water Supply Index

** Used to forecast and calculation of San Joaquin Valley 60-20-20 Water Supply Index

† In use for B120 point forecasts as of April1st, 2019

Table 11: Fitted fixed effects panel regression models with the outcome April 1st to July 31st basin release volume (N=14, T=34; n =416).

Dependent variable: Basin Release
Method :Panel (Fixed Effectts)

Sample: 1985-2018
Cross-sections included: 14
Total Panel (unbalanced) observations: 416
Standard errors robust to heteroskedasticity

Variable	Coefficent		Robust Standard Error	t-Statistic	p value
April 1st Forecast	(β_0)	0.648	0.059	11.032	0.000
March 31st Avalible Reservoir Storage Space	(β_1)	-0.601	0.166	-3.627	0.003
April 1st Forecast Uncertanty	(β_3)	-0.698	0.136	-5.120	0.000
Forecast x Uncertainty Interaction	(γ_1)	0.182	0.038	4.756	0.000

Weighted Statistics	
R-squared	0.709
Adjusted R-squared	0.696
TSS / ESS	803.085
RSS	38.976
Wald F	617.883
p value (Wald F)	0.000

Table 12: Individual effects for the 14 basins in the fixed effect panel data model.

Basin	Individual Fixed Effect (μ_i)	Robust Standard Error	t-Statistic	p-value
Sacramento	2.395	0.371	6.457	0.000
Feather	0.732	0.360	2.035	0.063
Yuba	0.580	0.331	1.755	0.103
American	0.931	0.341	2.730	0.017
Cosumnes	0.204	0.299	0.684	0.506
Mokolumne	0.463	0.315	1.468	0.166
Stanislaus	0.265	0.338	0.783	0.448
Tuolumne	0.235	0.341	0.691	0.502
Merced	0.602	0.319	1.885	0.082
San Joaquin	0.750	0.339	2.212	0.045
Kings	1.028	0.344	2.985	0.011
Kawea	0.653	0.417	1.565	0.142
Tule	0.443	0.402	1.101	0.291
Kern	0.680	0.441	1.541	0.147

Table 13: Reservoirs for each basin and their characteristics.

Basin Attributes			Individual Reservoir Attributes						
River Basin	Forecast and Release Point	Basin Storage Capacity (Km^3)	CDEC	Name	Operator	Storage Capacity (Km^3)	Elevation (meters)	Longitude	Latitude
Sacramento	SBB	6.1	IRC	Iron Canyon	PGE	0.03	815	-121.981	41.045
			WHI	Whiskeytown	USBR	0.30	369	-122.537	40.598
			MCO	McCloud	PGE	0.04	823	-122.074	41.135
			BIT	Britton	PGE	0.05	836	-121.676	41.022
			PT7	Pit No. 7	PGE	0.04	316	-122.015	40.843
			PT6	Pit No. 6	PGE	0.02	10	-122.001	40.842
			SHA	Shasta	USBR	5.61	325	-122.42	40.718
			KES	Keswick	USBR	0.03	182	-122.4457	40.6121
Feather	FTO	6.7	ORO	Oroville	CADWR	4.36	274	-121.493	39.54
			ALM	Almanor	PGE	1.61	1402	-121.173	40.218
			BCL	Bucks	PGE	0.13	1575	-121.202	39.897
			LGV	Little Grass Valley	SFWP	0.11	1538	-121.019	39.724
			DAV	Davis	CADWR	0.10	1758	-120.467	39.883
			SLC	Sly Creek	SFWP	0.08	1078	-121.116	39.584
			FRD	Frenchman	CADWR	0.07	1682	-120.183	39.883
			BTV	Butt Valley	PGE	0.06	1263	-121.186	40.187
			MMW	Mountain Meadows	PGE	0.03	1538	-121.017	40.283
			ANT	Antelope	CADWR	0.03	1512	-120.607	40.18
			TMT	Thermalito Total	CADWR	0.10	43	-121.638	39.458
Yuba	YRS	1.85	BUL	New Bullards Bar	YCWA	1.19	599	-121.14	39.393
			BWS	Lake Spaulding	YBDS	0.18	1643	-120.658	39.441
			ENG	Englebright	USACE	0.09	168	-121.267	39.239
			JCK	Jackson Meadows	YBDS	0.09	1840	-120.552	39.509
			SFL	Scotts Flat	YBDS	0.06	940	-120.917	39.274
			FRL	French Lake	YBDS	0.02	2031	-120.544	39.44
			RLL	Rollins	YBDS	0.08	663	-120.948	39.135
			BWN	Bowman	YBDS	0.08	1697	-120.652	39.45
American	AMF	2.17	FOL	Folsom	USBR	1.21	142	-121.183	38.683
			NAT	Natoma	USBR	0.01	40	-121.183	38.65
			UNV	Union Valley	SMUD	0.33	1484	-120.438	38.864
			HHL	Hell Hole	PCWA	0.26	1411	-120.407	39.058
			FMD	French Meadows	PCWA	0.17	1604	-120.468	39.112
			LON	Loon	SMUD	0.09	1954	-120.323	38.983
			ICH	Ice House	SMUD	0.05	1662	-120.359	38.824
			CPL	Caples	EDID	0.03	2377	-120.048	38.707
			EDN	Stumpy Meadows	GDPUD	0.02	1302	-120.603	38.903
			SLB	Slab Creek	SMUD	0.02	564	-120.6994	38.7725
			SIV	Silver	EDID	0.01	2174	-120.121	38.669
			LVY	Lake Valley	YBDS	0.01	1785	-120.583	39.3
Cosumnes	CSN	0.06	JNK	Jenkinson	EDID	0.05	1061	-120.56	38.713
Mokolumne	MKM	1.07	CMN	Camanche	EBMUD	0.51	72	-121.021	38.225
			PAR	Pardee	EBMUD	0.26	173	-120.85	38.25
			SLS	Salt Springs	PGE	0.18	1128	-120.216	38.497
			LWB	Lower Bear	PGE	0.06	1775	-120.257	38.538
Stanislaus	SNS	3.5	NML	New Melones	USBR	2.96	346	-120.525	37.9481
			SPM	New Spicer Meadows	NCPA	0.23	1986	-119.998	38.393
			BRD	Beardsley	TriDam	0.12	1036	-120.075	38.203
			TUL	Tulloch	TriDam	0.08	157	-120.603	37.875
			DON	Donnells	TriDam	0.08	1474	-119.96	38.329
			SWB	Pinecrest	PGE	0.02	1713	-119.983	38.2
			RLF	Relief	PGE	0.02	2201	-119.733	38.267
			LYS	Lyons	PGE	0.01	1260	-120.166	38.094
Tuolumne	TLG	3.44	DNP	New Don Pedro	TID	2.50	253	-120.421	37.702
			HTH	Hetch Hetchy	SF	0.44	1180	-119.783	37.95
			CHY	Cherry	SF	0.33	1437	-119.913	37.976
			TLC	Turlock	TID	0.06	74	-120.583	37.6
			MDO	Modesto	MID	0.04	64	-120.675	37.657
			ENR	Eleanor	SF	0.04	1420	-119.88	37.974
Merced	MRC	1.28	MCR	McClure	MID	1.26	264	-120.267	37.584
			MCS	McSwain	MID	0.01	130	-120.309	37.52
San Joaquin	SJF	1.42	MIL	Millerton	USBR	0.64	177	-119.705	37.001
			SHV	Shaver	SCE	0.17	1637	-119.302	37.145
			TAE	Thomas A. Edison	SCE	0.15	2329	-118.987	37.37
			MPL	Mammoth Pool	SCE	0.15	1015	-119.328	37.329
			HNT	Huntington	SCE	0.11	2134	-119.221	37.228
			FLR	Florence	SCE	0.08	2233	-118.967	37.267
			CNV	Crane Valley	PGE	0.06	1030	-119.517	37.283
			RDN	Redinger	SCE	0.04	428	-119.449	37.145
Kings	KGF	1.53	PNF	Pine Flat	USACE	1.23	296	-119.325	36.833
			CTG	Helms Creek	PGE	0.15	2497	-118.967	37.072
			WSN	Wishon	PGE	0.16	1999	-118.969	37.005
Kawea	KWT	0.23	TRM	Kaweah	USACE	0.23	229	-119.002	36.415
Tule	SCC	0.1	SCC	Success	USACE	0.10	211	-118.922	36.061
Kern	KRI	0.7	ISB	Isabella	USACE	0.70	803	-118.473	35.646

3.8 Figures

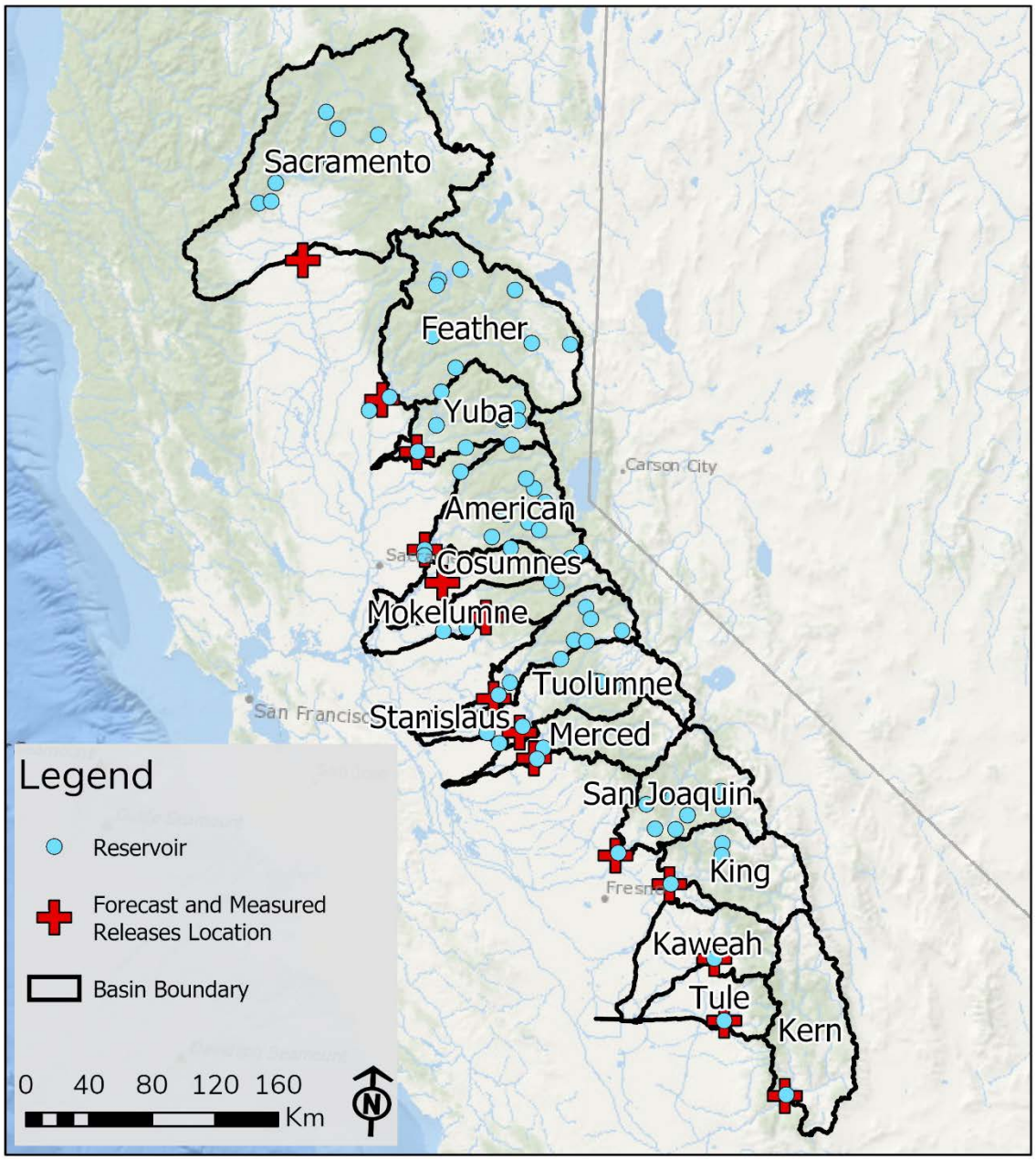

Figure 18: Study location. The 14 basins draining the west side of the Sierra Nevada. Blue dots are reservoirs used in the study and red marks are the location in each basin where releases are measured and the forecasts are produced.

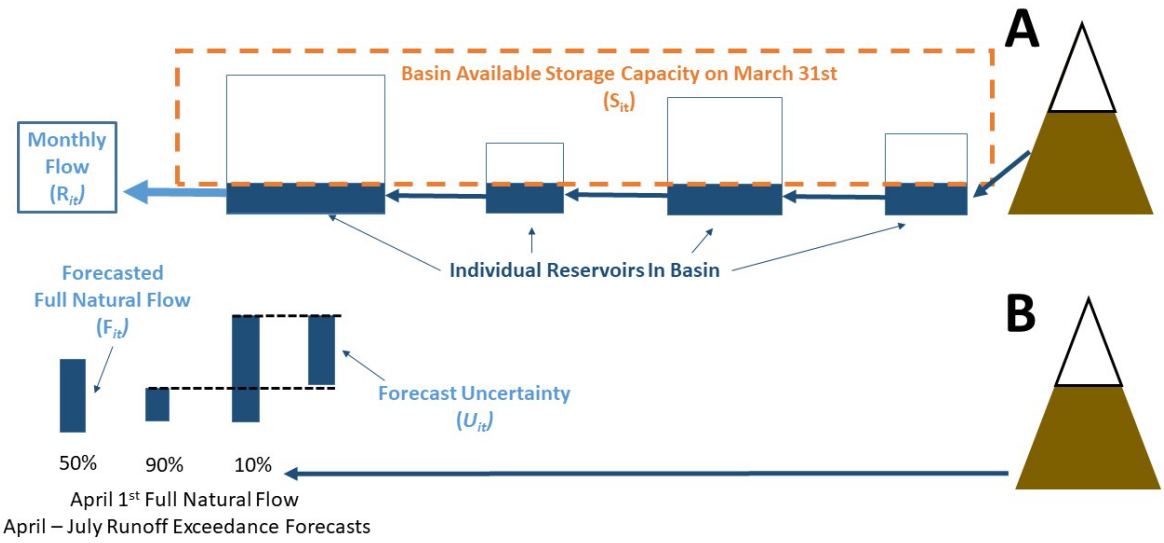

Figure 19. Conceptual model of variables used for each basin in the fixed effects panel data model. The top row (A) shows the measured hydrologic values from the basin and how they are used in the model. March 31st available space in individual reservoir are integrated into an overall measure of available space in the basin (S_{it}). The measured flow out of the basin each month is used as the measure of releases from the basin due to water management decisions (R_{it}). The second row (B) shows the forecasted values that are used in the panel data model. The 50% exceedance forecast on April 1st of April-July Full Natural Flow is the forecasted flow value in the model (F_{it}) and the spread between the 90% and the 10% exceedance values is used as the measure of forecast uncertainty with a wider spread representative of more uncertainty in upcoming flows (U_{it}).

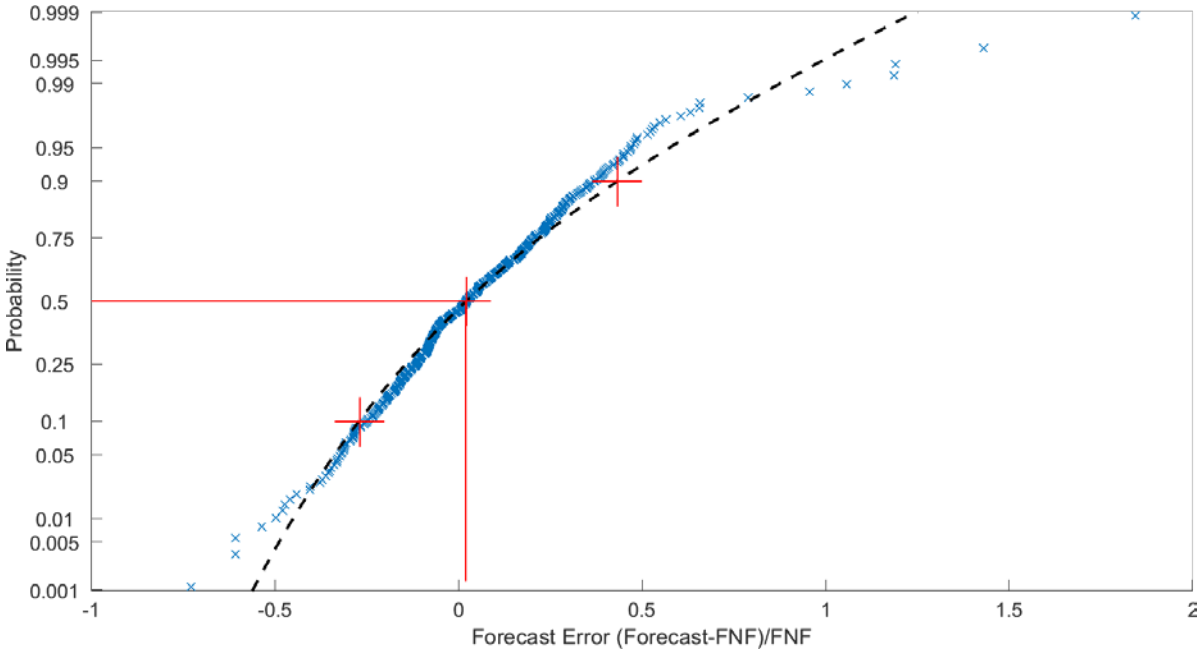

Figure 20: Errors distribution for historical April 1st Bulletin 120 forecasts of April-July full natural flows (FNF) in Sierra Nevada rivers (1985-2018, n = 442). The red lines indicate the median error, near 0%. The red crosses indicate the errors at the 10th

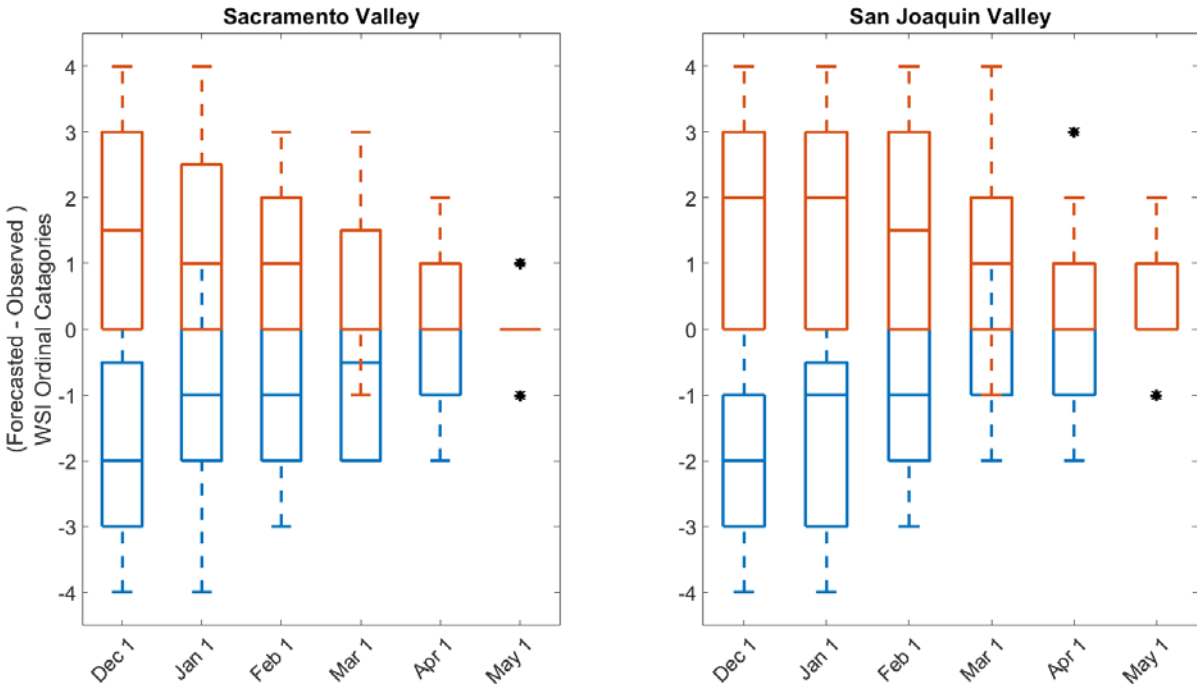

Figure 21: Official forecast uncertainty for the Water Supply Indices for the Sacramento and San Joaquin valley.

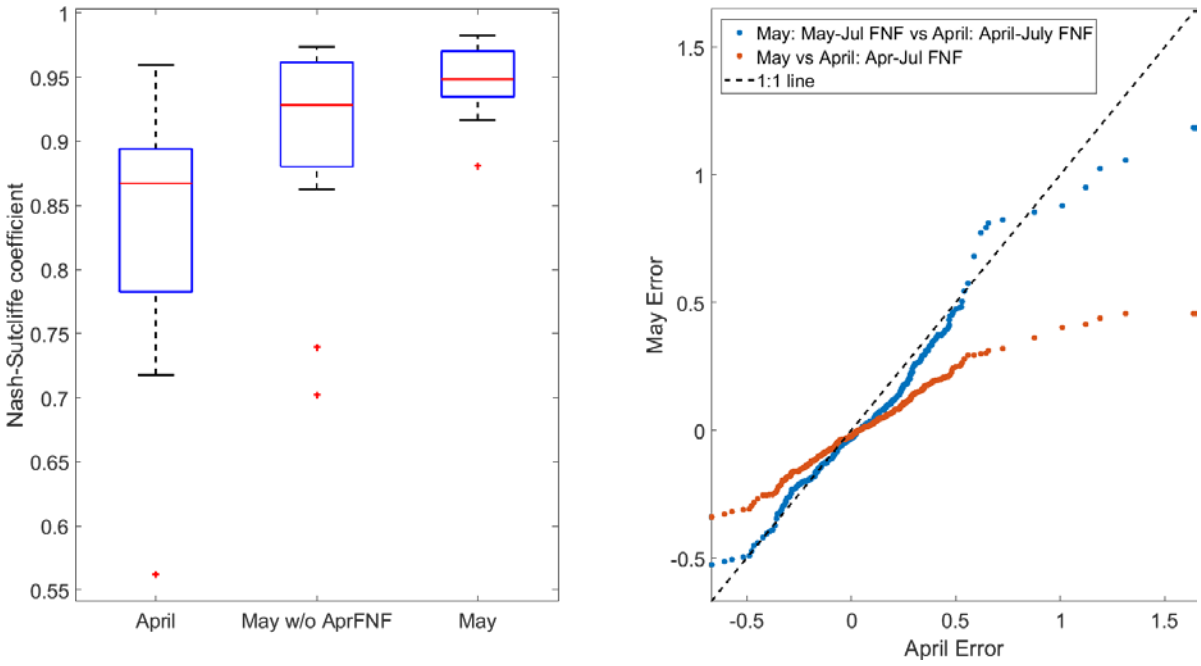

Figure 22: Perceived improvement in April – July runoff forecasts during the melt season. Left graph is boxplots of the Nash-Sutcliffe coefficient for the 14 basins. Right graph is Q-Q plots for comparing the April 1st April-July FNF forecast to both the May 1st April-July FNF forecast(red points) and the May 1st May-July FNF forecast (blue points). The error distribution of the May 1st forecast of May-July runoff is the same as the error distribution of the April 1st forecast of April – July runoff. There is no improvement in forecasting skill on May 1st compared to April 1st.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Susstrunk, S. (2012). SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 34*, 2274-2281. https://doi.org/10.1109/Tpami.2012.120

Álvarez, I. C., Barbero, J., & Zofío, J. L. (2017). A panel data toolbox for MATLAB. *Journal of Statistical Software, 76*, 1-27. https://doi.org/10.18637/jss.v076.i06

Andreadis, K. M., & Lettenmaier, D. P. (2006). Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources, 29*, 872-886. https://doi.org/10.1016/j.advwatres.2005.08.004

Bair, E. H., Rittger, K., Davis, R. E., Painter, T. H., & Dozier, J. (2016). Validating reconstruction of snow water equivalent in California's Sierra Nevada using measurements from the NASA Airborne Snow Observatory. *Water Resources Research, 52*, 8437-8460. https://doi.org/10.1002/2016WR018704

Bair, E. H., Rittger, K., Skiles, S. M., & Dozier, J. (2019). An examination of snow albedo estimates from MODIS and their impact on snow water equivalent reconstruction. *Water Resources Research, 55*. https://doi.org/10.1029/2019WR024810

Baldridge, A. M., Hook, S. J., Grove, C. I., & Rivera, G. (2009). The ASTER spectral library version 2.0. *Remote Sensing of Environment, 113*, 711-715. https://doi.org/10.1016/j.rse.2008.11.007

Barnard, J. M. (1989). *The value of inflow forecasting in the operation of a hydroelectric reservoir.* (Master of Applied Science), University of British Columbia, Vancouver, BC.

Barsi, J. A., Lee, K., Kvaran, G., Markham, B. L., & Pedelty, J. A. (2014). The spectral response of the Landsat-8 Operational Land Imager. *Remote Sensing, 6*, 10232-10251. https://doi.org/10.3390/rs61010232

Bianconi, F., & Fernandez, A. (2007). Evaluation of the effects of Gabor filter parameters on texture classification. *Pattern Recognition, 40*, 3325-3335. https://doi.org/10.1016/j.patcog.2007.04.023

Bohren, C. F., & Beschta, R. L. (1979). Snowpack albedo and snow density. *Cold Regions Science and Technology, 1*, 47-50. https://doi.org/10.1016/0165-232X(79)90018-1

Bohren, C. F., & Huffman, D. R. (2007). *Absorption and Scattering of Light by Small Particles*. New York: John Wiley and Sons. https://doi.org/10.1002/9783527618156

Burt, P., & Adelson, E. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications, 31*, 532-540. https://doi.org/10.1109/TCOM.1983.1095851

CADWR. (2018). *Flood-MAR: Using Flood Water for Managed Aquifer Recharge to Support Sustainable Water Resources*. California Department of Water Resources.

California Cooperative Snow Survey. (1930-2019). *Water Conditions in California Bulletin 120*. Sacramento, CA: California Department of Water Resources Retrieved from https://cdec.water.ca.gov/snow/bulletin120/.
California Energy Commission. (2018). Total System Electric Generation: 2017. *California Energy Almanac.* Retrieved from https://www.energy.ca.gov/almanac/electricity_data/total_system_power.html
California Natural Resources Agency. (2013). *Final California Water Plan Update 2013*.
Chai, D., Newsam, S., Zhang, H. K., Qiu, Y., & Huang, J. (2019). Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks. *Remote Sensing of Environment, 225*, 307-316. https://doi.org/10.1016/j.rse.2019.03.007
Church, J. E. (1914). Recent studies of snow in the United States. *Quarterly Journal of the Royal Meteorological Society, 40*, 43-52. https://doi.org/10.1002/qj.49704016905
Church, J. E. (1933). Snow surveying: Its principles and possibilities. *Geographical Review, 23*, 529-563. https://doi.org/10.2307/209242
Cox, L. M., Bartee, D., Crook, A., Farnes, P. E., & Smith, J. L. (1978). *The care and feeding of snow pillows*. Presented at the 46th Western Snow Conference, Otter Rock, Oregon. https://westernsnowconference.org/node/1109
Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol 1, Proceedings*, 886-893. Retrieved from <Go to ISI>://WOS:000230923100119
Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial-frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A, 2*, 1160-1169. https://doi.org/10.1364/Josaa.2.001160
Dettinger, M. D., Ralph, F. M., Das, T., Neiman, P. J., & Cayan, D. R. (2011). Atmospheric Rivers, Floods and the Water Resources of California. *Water, 3*, 445-478. https://doi.org/10.3390/w3020445
Dozier, J. (1989). Spectral signature of alpine snow cover from the Landsat Thematic Mapper. *Remote Sensing of Environment, 28*, 9-22. https://doi.org/10.1016/0034-4257(89)90101-6
Dozier, J., & Frew, J. (1990). Rapid calculation of terrain parameters for radiation modeling from digital elevation data. *IEEE Transactions on Geoscience and Remote Sensing, 28*, 963-969. https://doi.org/10.1109/36.58986
Dozier, J., & Painter, T. H. (2004). Multispectral and hyperspectral remote sensing of alpine snow properties. *Annual Review of Earth and Planetary Sciences, 32*, 465-494. https://doi.org/10.1146/annurev.earth.32.101802.120404
Dozier, J. (2011). Mountain hydrology, snow color, and the fourth paradigm. *Eos, 92*, 373-374. https://doi.org/10.1029/2011eo430001
Dozier, J., Bair, E. H., & Davis, R. E. (2016). Estimating the spatial distribution of snow water equivalent in the world's mountains. *WIREs Water, 3*, 461-474. https://doi.org/10.1002/wat2.1140

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., . . . Alsdorf, D. (2007). The Shuttle Radar Topography Mission. *Reviews of Geophysics, 45*, RG2004. https://doi.org/10.1029/2005RG000183

Fassnacht, S. R., Dressler, K. A., & Bales, R. C. (2003). Snow water equivalent interpolation for the Colorado River Basin from snow telemetry (SNOTEL) data. *Water Resources Research, 39*, 1208. https://doi.org/10.1029/2002WR001512

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A, 4*, 2379-2394. https://doi.org/10.1364/Josaa.4.002379

Foga, S., Scaramuzza, P. L., Guo, S., Zhu, Z., Dilley Jr, R. D., Beckmann, T., . . . Laue, B. (2017). Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sensing of Environment, 194*, 379-390. https://doi.org/10.1016/j.rse.2017.03.026

Frew, J. E., & Dozier, J. (2012). Environmental informatics. *Annual Review of Environment and Resources, 37*, 449-472. https://doi.org/10.1146/annurev-environ-042711-121244

Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., . . . Schaaf, C. (2002). Global land cover mapping from MODIS: algorithms and early results. *Remote Sensing of Environment, 83*, 287-302. https://doi.org/10.1016/S0034-4257(02)00078-0

Gabor, D. (1946). Theory of communication. Part 1: The analysis of information. *Journal of the Institution of Electrical Engineers – Part III: Radio and Communication Engineering, 93*, 429-441. https://doi.org/10.1049/ji-3-2.1946.0074

Gafurov, A., & Bárdossy, A. (2009). Cloud removal methodology from MODIS snow cover product. *Hydrology and Earth System Science, 13*, 1361-1373. https://doi.org/10.5194/hess-13-1361-2009

Gomez-Chova, L., Camps-Valls, G., Calpe-Maravilla, J., Guanter, L., & Moreno, J. (2007). Cloud-screening algorithm for ENVISAT/MERIS multispectral images. *IEEE Transactions on Geoscience and Remote Sensing, 45*, 4105-4118. https://doi.org/10.1109/TGRS.2007.905312

Guan, B., Waliser, D. E., Molotch, N. P., Fetzer, E. J., & Neiman, P. J. (2012). Does the Madden-Julian Oscillation influence wintertime atmospheric rivers and snowpack in the Sierra Nevada? *Monthly Weather Review, 140*, 325-342. https://doi.org/10.1175/Mwr-D-11-00087.1

Gueymard, C. A. (2005). Interdisciplinary applications of a versatile spectral solar irradiance model: A review. *Energy, 30*, 1551-1576. https://doi.org/10.1016/j.energy.2004.04.032

Hale, G. M., & Querry, M. R. (1973). Optical constants of water in the 200-nm to 200-μm wavelength region. *Applied Optics, 12*, 555-563. https://doi.org/10.1364/AO.12.000555

Hall, D. K., & Riggs, G. A. (2007). Accuracy assessment of the MODIS snow products. *Hydrological Processes, 21*, 1534-1547. https://doi.org/10.1002/hyp.6715

Harrison, B., & Bales, R. (2016). Skill assessment of water supply forecasts for Western Sierra Nevada watersheds. *Journal of Hydrologic Engineering, 21*. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001327

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Berlin: Springer. https://doi.org/10.1007/978-0-387-84858-7

Helfrich, S. R., McNamara, D., Ramsay, B. H., Baldwin, T., & Kasheta, T. (2007). Enhancements to, and forthcoming developments in the Interactive Multisensor Snow and Ice Mapping System (IMS). *Hydrological Processes, 21*, 1576-1586. https://doi.org/10.1002/hyp.6720

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in 2 nonstriate visual areas (18 and 19) of cat. *Journal of Neurophysiology, 28*, 229-289. https://doi.org/10.1152/jn.1965.28.2.229

Huber, A. L., & Robertson, D. C. (1982). *Regression models in water supply forecasting*. Presented at the 50th Western Snow Conference, Reno, Nevada. https://westernsnowconference.org/node/922

Hughes, M. J., & Hayes, D. J. (2014). Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and spatial post-processing. *Remote Sensing, 6*, 4907-4926. https://doi.org/10.3390/rs6064907

Irish, R. R., Barker, J. L., Goward, S. N., & Arvidson, T. (2006). Characterization of the Landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. *Photogrammetric Engineering and Remote Sensing, 72*, 1179-1188.

Jain, A. K., Ratha, N. K., & Lakshmanan, S. (1997). Object detection using Gabor filters. *Pattern Recognition, 30*, 295-309. https://doi.org/10.1016/S0031-3203(96)00068-4

Jasperse, J., Ralph, M., Anderson, M., Brekke, L. D., Dillabough, M., Dettinger, M. D., . . . Webb, R. H. (2017). *Preliminary viability assessment of Lake Mendocino forecast informed reservoir operations*. La Jolla, CA: Center for Western Weather and Water Extremes. Retrieved from http://pubs.er.usgs.gov/publication/70192184

JPL AVIRIS Team. (2018). AVIRIS Next Generation. Retrieved from https://avirisng.jpl.nasa.gov/

Justice, C. O., Vermote, E., Townshend, J. R. G., Defries, R., Roy, D. P., Hall, D. K., . . . Barnsley, M. J. (1998). The Moderate Resolution Imaging Spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing, 36*, 1228-1249. https://doi.org/10.1109/36.701075

Karagulle, D., Frye, C., Sayre, R., Breyer, S., Aniello, P., Vaughan, R., & Wright, D. (2017). Modeling global Hammond landform regions from 250-m elevation data. *Transactions in GIS, 21*, 1040-1060. https://doi.org/10.1111/tgis.12265

Kou, L., Labrie, D., & Chylek, P. (1993). Refractive indices of water and ice in the 0.65- to 2.5-μm spectral range. *Applied Optics, 32*, 3531-3540. https://doi.org/10.1364/AO.32.003531

Kruse, F. A., Lefkoff, A. B., Boardman, J. W., Heidebrecht, K. B., Shapiro, A. T., Barloon, P. J., & Goetz, A. F. H. (1993). The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. *Remote*

Sensing of Environment, 44, 145-163. https://doi.org/10.1016/0034-4257(93)90013-N
Krzysztofowicz, R. (1986). Optimal water supply planning based on seasonal runoff forecasts. *Water Resources Research, 22*, 313-321. https://doi.org/10.1029/WR022i003p00313
Kumar, A., & Pang, G. K. H. (2002). Defect detection in textured materials using Gabor filters. *IEEE Transactions on Industry Applications, 38*, 425-440. https://doi.org/10.1109/28.993164
Lamei, N., Hutchison, K. D., Crawford, M. M., & Khazenie, N. (1994). Cloud-type discrimination via multispectral textural analysis. *Optical Engineering, 33*. https://doi.org/10.1117/12.166920
Li, Z. W., Shen, H. F., Li, H. F., Xia, G. S., Gamba, P., & Zhang, L. P. (2017). Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery. *Remote Sensing of Environment, 191*, 342-358. https://doi.org/10.1016/j.rse.2017.01.026
Li, Z. W., Shen, H. F., Cheng, Q., Liu, Y. H., You, S. C., & He, Z. Y. (2019). Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors. *ISPRS Journal of Photogrammetry and Remote Sensing, 150*, 197-212. https://doi.org/10.1016/j.isprsjprs.2019.02.017
Mankin, J. S., Viviroli, D., Singh, D., Hoekstra, A. Y., & Diffenbaugh, N. S. (2015). The potential for snow to supply human water demand in the present and future. *Environmental Research Letters, 10*, 114016. https://doi.org/10.1088/1748-9326/10/11/114016
Meerdink, S. K., Hook, S. J., Roberts, D. A., & Abbott, E. A. (2019). The ECOSTRESS spectral library version 1.0. *Remote Sensing of Environment, 230*, 111196. https://doi.org/10.1016/j.rse.2019.05.015
NASA MODIS Team. (2003). MODIS Technical Specifications. Retrieved from https://modis.gsfc.nasa.gov/about/specifications.php
NASA/USGS Landsat 8 Team. (2014). Spectral response of the Operational Land Imager in-band, band-average relative spectral response. Retrieved from https://landsat.gsfc.nasa.gov/preliminary-spectral-response-of-the-operational-land-imager-in-band-band-average-relative-spectral-response/
National Academies of Sciences, Engineering, & Medicine. (2018). *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space*. Washington, DC: National Academies Press. https://doi.org/10.17226/24938
Nussenzveig, H. M., & Wiscombe, W. J. (1980). Efficiency factors in Mie scattering. *Physical Review Letters, 45*, 1490-1494. https://doi.org/10.1103/PhysRevLett.45.1490
Odegard, H. L., Eidsvik, J., & Fleten, S. E. (2019). Value of information analysis of snow measurements for the scheduling of hydropower production. *Energy Systems-Optimization Modeling Simulation and Economic Aspects, 10*, 1-19. https://doi.org/10.1007/s12667-017-0267-3
Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*. Berlin: Springer Verlag. https://doi.org/10.1007/978-3-540-76917-0
Pagano, T., Garen, D., & Sorooshian, S. (2004). Evaluation of official western US seasonal water supply outlooks, 1922-2002. *Journal of Hydrometeorology, 5*,

896-909. https://doi.org/10.1175/1525-7541(2004)005<0896:Eoowus>2.0.Co;2

Painter, T. H., Berisford, D. F., Boardman, J. W., Bormann, K. J., Deems, J. S., Gehrke, F., . . . Winstral, A. (2016). The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. *Remote Sensing of Environment, 184*, 139-152. https://doi.org/10.1016/j.rse.2016.06.018

Painter, T. H., Skiles, S. M., Deems, J. S., Brandt, W. T., & Dozier, J. (2018). Variation in rising limb of Colorado River snowmelt runoff hydrograph controlled by dust radiative forcing in snow. *Geophysical Research Letters, 45*, 797–808. https://doi.org/10.1002/2017GL075826

Palazzo, J., Liu, O. R., Stillinger, T., Song, R., Wang, Y., Hiroyasu, E. H. T., . . . Tague, C. (2017). Urban responses to restrictive conservation policy during drought. *Water Resources Research, 53*, 4459-4475. https://doi.org/10.1002/2016wr020136

Piechota, T. C., & Dracup, J. A. (1996). Drought and regional hydrologic variation in the United States: Associations with the El Niño Southern Oscillation. *Water Resources Research, 32*, 1359-1373. https://doi.org/10.1029/96wr00353

Platnick, S., King, M. D., Ackerman, S. A., Menzel, W. P., Baum, B. A., Riedi, J. C., & Frey, R. A. (2003). The MODIS cloud products: Algorithms and examples from Terra. *IEEE Transactions on Geoscience and Remote Sensing, 41*, 459-473. https://doi.org/10.1109/Tgrs.2002.808301

Pollen, D. A., & Ronner, S. F. (1983). Visual cortical-neurons as localized spatial-frequency filters. *IEEE Transactions on Systems Man and Cybernetics, 13*, 907-916. https://doi.org/10.1109/Tsmc.1983.6313086

Qiu, S., He, B. B., Zhu, Z., Liao, Z. M., & Quan, X. W. (2017). Improving Fmask cloud and cloud shadow detection in mountainous area for Landsats 4-8 images. *Remote Sensing of Environment, 199*, 107-119. https://doi.org/10.1016/j.rse.2017.07.002

Rheinheimer, D. E., Bales, R. C., Oroza, C. A., Lund, J. R., & Viers, J. H. (2016). Valuing year-to-go hydrologic forecast improvements for a peaking hydropower system in the Sierra Nevada. *Water Resources Research, 52*, 3815-3828. https://doi.org/10.1002/2015wr018295

Rittger, K., Bair, E. H., Kahl, A., & Dozier, J. (2016). Spatial estimates of snow water equivalent from reconstruction. *Advances in Water Resources, 94*, 345-363. https://doi.org/10.1016/j.advwatres.2016.05.015

Rittger, K., Raleigh, M. S., Hill, A. F., Painter, T. H., & Dozier, J. (2019). Canopy adjustment and improved cloud detection for remotely sensed snow mapping. *Water Resources Research*, in review.

Roos, M. (2003). *California's Cooperative Snow Surveys Program*. Presented at the 49th Annual Meeting of California Snow Survey Cooperators, Folsom, CA. http://cdec.water.ca.gov/snow/meeting/2003/Calif_Coop_Snow_Program.pdf

Rouse, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. (1974). *Monitoring vegetation systems in the Great Plains with ERTS.* Paper presented at the 3rd Earth Resource Technology Satellite (ERTS) Symposium. Retrieved from https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19740022592.pdf

Roy, D. P., Wulder, M. A., Loveland, T. R., Woodcock, C. E., Allen, R. G., Anderson, M. C., . . . Zhu, Z. (2014). Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment, 145*, 154-172. https://doi.org/10.1016/j.rse.2014.02.001

Selkowitz, D. J., Forster, R. R., & Caldwell, M. K. (2014). Prevalence of pure versus mixed snow cover pixels across spatial resolutions in alpine environments. *Remote Sensing, 6*, 12478-12508. https://doi.org/10.3390/rs61212478

Skiles, S. M., Painter, T. H., Deems, J. S., Bryant, A. C., & Landry, C. C. (2012). Dust radiative forcing in snow of the Upper Colorado River Basin: 2. Interannual variability in radiative forcing and snowmelt rates. *Water Resources Research, 48*, W07522. https://doi.org/10.1029/2012wr011986

Space Applications Centre. (2017). *Spectrum of India*. Bangalore: Indian Space Research Organisation.

Water Right Decision 1485, State Water Resources Control Board Decision 1485 C.F.R. (1978).

Revised Water Right Decision 1641, State Water Resources Control Board Decision 1641 C.F.R. (1999).

Stillinger, T., & Collar, N. M. (2019). *Snow-Cloud Validation Masks for Multispectral Satellite Data*. Version 1. https://doi.org/10.5281/zenodo.3240937

Stillinger, T., Roberts, D. A., Collar, N. M., & Dozier, J. (2019). Cloud masking for Landsat 8 and MODIS Terra over snow-covered terrain: Error analysis and spectral similarity between snow and cloud. *Water Resources Research, 55*, 6169-6184. https://doi.org/10.1029/2019WR024932

Strauss, F. A. (1954). Forecasting water supply through snow surveys. *Journal American Water Works Association, 46*, 853-863. Retrieved from http://www.jstor.org/stable/41253116

Stubenrauch, C. J., Rossow, W. B., Kinne, S., Ackerman, S., Cesana, G., Chepfer, H., . . . Zhao, G. (2013). Assessment of global cloud datasets from satellites: Project and database initiated by the GEWEX Radiation Panel. *Bulletin of the American Meteorological Society, 94*, 1031-1049. https://doi.org/10.1175/Bams-D-12-00117.1

Sturm, M., Goldstein, M. A., & Parr, C. (2017). Water and life from snow: A trillion dollar science question. *Water Resources Research, 53*, 3534-3544. https://doi.org/10.1002/2017wr020840

Tran, H., Nguyen, P., Ombadi, M., Hsu, K.-l., Sorooshian, S., & Qing, X. (2019). A cloud-free MODIS snow cover dataset for the contiguous United States from 2000 to 2017. *Scientific Data, 6*, 180300. https://doi.org/10.1038/sdata.2018.300

U.S. Bureau of Economic Analysis. (2018). Real Total Gross Domestic Product for California [CARGSP]. Federal Reserve Bank of St. Louis. Retrieved from https://fred.stlouisfed.org/series/CARGSP

U.S. Geological Survey. (2016). *L8 Biome Cloud Validation Masks*. https://doi.org/10.5066/F7FB5146

U.S. Geological Survey. (2017). *Landsat 8 Level-1 Data Format Control Book, Version 11.0*: U.S. Geological Survey. Retrieved from https://www.usgs.gov/media/files/landsat-8-level-1-data-format-control-book

Warren, S. G., & Wiscombe, W. J. (1980). A model for the spectral albedo of snow, II, Snow containing atmospheric aerosols. *Journal of the Atmospheric Sciences, 37*, 2734-2745. https://doi.org/10.1175/1520-0469(1980)037<2734:AMFTSA>2.0.CO;2

Warren, S. G. (1982). Optical properties of snow. *Reviews of Geophysics, 20*, 67-89. https://doi.org/10.1029/RG020i001p00067

Warren, S. G., & Brandt, R. E. (2008). Optical constants of ice from the ultraviolet to the microwave: A revised compilation. *Journal of Geophysical Research: Atmospheres, 113*, D14220. https://doi.org/10.1029/2007JD009744

Wester, P., Mishra, A., Mukherji, A., & Shrestha, A. B. (Eds.). (2019). *The Hindu Kush Himalaya Assessment: Mountains, Climate Change, Sustainability and People*. Kathmandu: ICIMOD. Retrieved from http://lib.icimod.org/record/34383.

Williams, I. N., & Patricola, C. M. (2018). Diversity of ENSO Events Unified by Convective Threshold Sea Surface Temperature: A Nonlinear ENSO Index. *Geophysical Research Letters, 45*, 9236-9244. https://doi.org/10.1029/2018gl079203

Wiscombe, W. J., & Warren, S. G. (1980). A model for the spectral albedo of snow, I, Pure snow. *Journal of the Atmospheric Sciences, 37*, 2712-2733. https://doi.org/10.1175/1520-0469(1980)037<2712:AMFTSA>2.0.CO;2

Wood, R., & Field, P. R. (2011). The distribution of cloud horizontal sizes. *Journal of Climate, 24*, 4800-4816. https://doi.org/10.1175/2011jcli4056.1

Wrzesien, M. L., Durand, M. T., Pavelsky, T. M., Kapnick, S. B., Zhang, Y., Guo, J., & Shum, C. K. (2018). A new estimate of North American mountain snow accumulation from regional climate model simulations. *Geophysical Research Letters, 45*, 1423-1432. https://doi.org/10.1002/2017GL076664

Zhang, H., Zhang, F., Zhang, G., Che, T., Yan, W., Ye, M., & Ma, N. (2019). Ground-based evaluation of MODIS snow cover product V6 across China: Implications for the selection of NDSI threshold. *Science of The Total Environment, 651*, 2712-2726. https://doi.org/10.1016/j.scitotenv.2018.10.128

Zhang, Z., Glaser, S., Bales, R., Conklin, M., Rice, R., & Marks, D. (2017). Insights into mountain precipitation and snowpack from a basin-scale wireless-sensor network. *Water Resources Research, 53*, 6626-6641. https://doi.org/10.1002/2016wr018825

Zhu, Z., & Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment, 118*, 83-94. https://doi.org/10.1016/j.rse.2011.10.028

Zhu, Z., & Woodcock, C. E. (2014a). Automated cloud, cloud shadow, and snow detection in multitemporal Landsat data: An algorithm designed specifically for monitoring land cover change. *Remote Sensing of Environment, 152*, 217-234. https://doi.org/10.1016/j.rse.2014.06.012

Zhu, Z., & Woodcock, C. E. (2014b). Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment, 144*, 152-171. https://doi.org/10.1016/j.rse.2014.01.011

Zhu, Z., Wang, S., & Woodcock, C. E. (2015). Improvement and expansion of the Fmask algorithm: cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and

Sentinel 2 images. *Remote Sensing of Environment, 159*, 269-277. https://doi.org/10.1016/j.rse.2014.12.014

Zi, Y., Xie, F. Y., & Jiang, Z. G. (2018). A Cloud Detection Method for Landsat 8 Images Based on PCANet. *Remote Sensing, 10*. https://doi.org/ARTN 877 10.3390/rs10060877